Genetic and Epigenetic Risk Factors for Invasive Lobular Breast Cancer

Medha Suman
Orchid id: 8432-7227

Submitted in total fulfilment of the requirements of the degree of

Doctor of Philosophy

April 2021

Department of Clinical Pathology

Faculty of Medicine, Dentistry and Health Sciences

The University of Melbourne.

Thesis abstract

Invasive lobular breast cancer (ILBC) is the second most common histological subtype of breast cancer and accounts for 10-15% of all cases. Loss of e-cadherin protein is a hallmark of ILBC and contributes to its characteristic discohesive morphology. In addition to distinct histological features, several subtype-specific molecular and clinical features have been described. However, ILBC remains understudied relative to other breast cancer subtypes, despite its frequency.

In this era of precision medicine, there is a growing interest in further refining breast cancer tumour subtyping by identifying additional discriminating molecular features. However, there is limited data to pursue this for ILBC as it is often not well-represented in study samples. For instance, the seminal work on breast cancer classification based on gene-expression levels by Perou et al., (2000) included only two ILBC cases. It is important to identify ways to refine the subtyping of ILBC tumours so that women with ILBC can benefit from a more precise treatment plan, prognosis and targeted therapy options.

The main objectives of this PhD project were: i) to examine the distinguishing methylation patterns between ILBC (n=151) and non-ILBC (n=341) tumours ii) to investigate the ILBC methylome to identify methylation signatures for prognostication (n=130) and iii) to subclassify ILBC into subgroups with increased homogeneity based on their genome-wide DNA methylation profiles (ILBC, n=151, non-ILBC=341) and to further characterise these subgroups by investigating their somatic mutational signatures (n=15).

Three subgroups of ILBC were defined via unsupervised cluster analysis of genome-wide DNA methylation measured using the Infinium HumanMethylation450K assay. Of these, Subgroup 1 was identified as the most distinct ILBC subgroup, characterised by a predominant hypomethylation across 27,675 CpGs compared with Subgroup 2 and across 13,067 CpGs compared with Subgroup 3. Subgroup 1 showed more similarity to the TNBC (non-ILBC) cases compared with the other two methylation-defined subgroups in terms of their genome-wide methylation pattern. Survival analysis showed that women with ILBC tumours in Subgroup 1 had the poorest overall survival when compared with women in Subgroup 2 (hazard ratio (HR): 0.59, 95% confidence interval (CI): 0.19-1.79) and Subgroup 3 (HR: 0.16, 95% CI: 0.03-0.88), after adjusting for age and year of diagnosis. Subgroup 3 had an enrichment for women who had a first-degree relative with a history of any cancer. Both Subgroup 2 and Subgroup 3 were enriched with women who had a female relative with a history of breast cancer. This suggests that women in Subgroup 2 and Subgroup 3 may be genetically or epigenetically predisposed to developing breast cancer.

The somatic genetic variant profiles of the ILBC DNA methylation-defined subgroups were further investigated by performing whole-exome sequencing (WES) on five ILBC tumours representing each of the three subgroups (n=15). The mismatch repair deficiency (MMRd) associated mutational signature SBS6 was the most frequently observed mutational signature in the ILBC tumours, detected in 12/15 (80%) cases. Microsatellite instability (MSI) was also predicted in 13/15 (87%) of the cases, including all 12 tumours with SBS6. Although distinct somatic (genetic) characteristics for tumours of individual subgroups were not observed, this research highlighted the potential role of MMRd in ILBC tumourigenesis and progression.

DNA methylation of ILBC was also investigated as a possible prognostic biomarker. The analysis revealed 2,771 variably methylated regions within the ILBC tumours (n=130). A pooled survival analysis of the study set and TCGA data identified *APC*, *TMEM101*, *HCG4P3* and *CELF2* promoter methylation as potential prognostic biomarkers for women with ILBC.

Comparing the DNA methylation profiles of ILBC (n=151) and non-ILBC (n=341) tumours, 13,763 genes and 8,456 intergenic regions showing statistically significant differences in DNA methylation (false discovery rate (fdr), *P*-value < 0.01) were identified. Gene set enrichment analysis revealed that the differentially methylated genes were found to be involved in biological pathways related to *metabolism of RNA* (R-HSA-8953854), *mRNA processing* (GO:0006397), *RNA splicing* (GO:0008380), *cell cycle* (R-HSA-1640170) and *DNA repair* (GO:0006281).

This study brings together several lines of evidence to indicate that distinct molecular features of ILBC can enable further subtyping, identify important features for targeted therapies (e.g., MMRd) and provide additional information for prognostication. This research identified Subgroup 1 as an important subgroup with similarities to TNBC and more aggressive clinical behaviour. Further investigation of samples from Subgroup 1 may identify additional important targets for precision medicine.

Declaration

The work in this thesis is original work carried out by the candidate unless otherwise indicated in the Preface.

Due acknowledgement has been made in the text to all other materials used.

The thesis is less than 100,000 words in length, exclusive of tables, figures, bibliography and appendices and compiles with the stipulations set out for the degree of Doctor of Philosophy by The University of Melbourne.

Medha Suman

Department of Clinical Pathology

The University of Melbourne

Parkville Victoria 3052

Australia.

Preface

DNA extraction from FFPE tumour tissue and the measurement of genome-wide DNA methylation for 402 breast cancer cases in this study was done by Dr Ee Ming Wong at the Genetic Epidemiology lab, Department of Clinical Pathology, The University of Melbourne. For the remaining 90 breast cancer cases, DNA methylation data was generated during this project by the candidate.

Work carried out in collaboration with others is as follows:

1. Statistical analyses of the association of variably methylated tumour DNA regions with overall survival for ILBC was performed by Dr Pierre-Antoine Dugué at Precision Medicine, School of Clinical Sciences at Monash Health, Monash University.

2. The bioinformatics pipeline used to process sequencing data and somatic variant calling was developed by Dr Jason Steen at Precision Medicine, School of Clinical Sciences at Monash Health, Monash University.

3. Pathology review of the breast tumours was conducted by Professor Catriona McLean at the Department of Anatomical Pathology, Alfred Hospital, Melbourne.

Research presented in this thesis was made possible only with the generous participation of women and their families in the studies MCCS, kConFab and ABCFR and their consent that allowed us to access to their pathology material and clinical data.

Acknowledgments

Undertaking this PhD has been a life-changing experience, and I do not see myself finishing this journey without acknowledging the contribution of many people for their guidance and support.

I would first like to thank my supervisors Professor Melissa C. Southey, Dr Tu Nguyen-Dumont and Dr Eric Ji-Hoon Joo whose expertise was invaluable throughout the entire project. Without their guidance, this PhD would not have been achievable. I am especially thankful to Professor Melissa Southey, who has been a strong support and a constant source of encouragement. She has been a great example of a leader and a person I look up to.

My huge thanks to Dr Pierre-Antoine Dugue for not only providing support with statistics but also providing me with invaluable advice as a mentor. I consider myself fortunate to have gotten an opportunity to collaborate with him that made it possible for me to get results from this project published. His inspiring approach towards research motivated me and prepared me in many ways for my future career too.

I thank every member of the Precision Medicine team who has always been supportive of each other, making our group an excellent place for research. I am truly grateful to Dr Ee Ming Wong, who trained me in the lab at the beginning of my PhD and contributed significantly to this project. Thanks to Dr Jason Steen for providing support with bioinformatics. Thanks to Ms Helen Tsimiklis and the whole biorepository team for

always being prompt and ready to provide the necessary support. My especial thanks to Fleur, who has been an emotional support for me in the lab when things did not go well.

I gratefully acknowledge the financial support provided by Professor Melissa Southey that allowed me to finish my candidature successfully. I would also like to recognise the Beaney Scholarship in Pathology that I received from the School of Biomedical Sciences, The University of Melbourne.

Lastly, I would like to say a heartfelt thanks to my mother and father for always believing in me and making me the person I am today. The values that they have instilled in me has prepared me to combat any challenge in life. I am thankful to the love and support of my dear husband; without his constant encouragement, I would not have embarked upon this journey in the first place. And my darling daughter for being such a good baby throughout the writing phase of my PhD and for making it possible for me to finish what I started.

Table of Contents

List of tables

List of figures

Publication arising from this thesis

1. **Clinical Epigenetics (Published 18 January 2021).**

   Association of variably methylated tumour DNA regions with overall survival for invasive lobular breast cancer.

   Suman M, Dugué P-A*, Wong EM, Joo JE, Hopper JL, Nguyen-Dumont T, . . . Southey MC. (2021). Association of variably methylated tumour DNA regions with overall survival for invasive lobular breast cancer. *Clinical epigenetics, 13*(1), 1-16.

   ** contributed equally to this work.*

Abstracts

1. **Human Genome Meeting: Perth, Australia 2020 - speaker selected for oral presentation from abstract.**

   Genome-wide Variably methylated tumour DNA regions and associations with overall survival in Invasive Lobular Breast Cancer.

   Medha Suman, Pierre-Antoine Dugué, Ee Ming Wong, JiHoon Eric Joo, John L. Hopper, Tu Nguyen-Dumont, Graham G. Giles, Roger L. Milne, Catriona McLean, Melissa C. Southey.

2. **Victorian Cancer Bioinformatics symposium: Melbourne, Australia 2019 - Poster presentation**

3. **Familial Aspects of Cancer Conference, Kingscliff, New South Wales, Australia 2019 – Poster presentation.**

   Invasive Lobular Breast Cancer: An integrated genetic and epigenetic approach to characterise Lobular tumours.

   Medha Suman, JiHoon Eric Joo, Tu Nguyen-Dumont, Jason Steen, Ee Ming Wong, Neil O'Callaghan, Melissa Yow, ABCFS, MCCS, kConFab, John L. Hopper, Graham G. Giles, Roger L. Milne, Melissa C. Southey.

4. **Lorne Cancer Conference, Lorne, Australia 2018 -Poster presentation.**
5. **Lorne Genome Conference, Lorne, Australia 2018 -Poster presentation.**

   Invasive Lobular Breast Cancer: Using tumour genome-wide DNA methylation for subtyping and aid in the identification of susceptibility genes.

   Medha Suman, JiHoon Eric Joo, Tu Nguyen-Dumont, Jason Steen, Ee Ming Wong, Neil O'Callaghan, Melissa Yow, ABCFS, MCCS, kConFab, John L. Hopper, Graham G. Giles, Roger L. Milne, Melissa C. Southey.

6. **Familial Aspects of Cancer Conference, Kingscliff, New South Wales, Australia 2018 – Poster presentation.**

   Invasive Lobular Breast Cancer: Using tumour genome-wide DNA methylation for subtyping and aid in the identification of susceptibility genes.

   Medha Suman, JiHoon Eric Joo, Tu Nguyen-Dumont, Jason Steen, Ee Ming Wong, Neil O'Callaghan, Melissa Yow, ABCFS, MCCS, kConFab, John L. Hopper, Graham G. Giles, Roger L. Milne, Melissa C. Southey.

7. **Familial Aspects of Cancer Conference, Kingscliff, New South Wales, Australia 2017 – Poster presentation.**

   Invasive Lobular Breast Cancer: Using tumour genome-wide DNA methylation for subtyping and aid in the identification of susceptibility genes.

   Medha Suman, JiHoon Eric Joo, Tu Nguyen-Dumont, Jason Steen, Ee Ming Wong, Neil O'Callaghan, Melissa Yow, ABCFS, MCCS, kConFab, John L. Hopper, Graham G. Giles, Roger L. Milne, Melissa C. Southey.

Abbreviations

ABCFR: Australian Breast Cancer Family Registry

AJCC: The American Joint Committee on Cancer

ASR: Age-standardised rate

CI: Confidence interval

CRE v2: Clinical research exome v2

DCIS: Ductal carcinoma in-situ

DMP: Differentially methylated position

DMR: Differentially methylated region

ER: Estrogen receptor

FFPE: Formalin-fixed paraffin-embedded

FNORM: Functional normalisation

FRR: Familial relative risk

GC: Guthrie Card

GWAS: Genome-wide association study

HER2: Human epidermal growth factor receptor 2

HM450K: HumanMethylation450K

HR: Hazard ratio

HRD: Homologous recombination deficiency

HRT: Hormone replacement therapy

ICD-O: International Classification of Diseases for Oncology

IDBC: Invasive ductal breast cancer

IHC: Immunohistochemistry

ILBC: Invasive lobular breast cancer

kConFab: The Kathleen Cuningham Foundation Consortium for Research into Familial Breast Cancer

KEGG : Kyoto Encyclopedia of Genes and Genomes

LCIS: Lobular carcinoma in-situ

LOH: Loss of heterozygosity

MAF: Minor allele frequency

MCCS: Melbourne Collaborative Cohort Study

MHT: Menopausal hormone therapy

MMRd: Mismatch repair deficiency

PI3K: Phosphatidylinositol 3-kinase

PMD: Percentage mammographic density

PR: Progesterone receptor

RR: Relative risk

SBS: Single base substitution

SNP: Single nucleotide polymorphism

SNV: Single nucleotide variant

SSNV: Somatic single nucleotide variant

TCGA: The Cancer Genome Atlas

TDLU: Terminal duct lobular unit

TMB: Tumour mutation burden

TNBC: Triple negative breast cancer

TSS: Transcription start site

UMI: Unique molecular index

VMR: Variably methylated region

WES: Whole-exome sequencing

Chapter 1 Literature Review

1.1 Human breast anatomy

Breast cancer is a heterogeneous disease where malignant (cancer) cells develop in the breast epithelial tissue (Harris *et al.*, 1992). Human breasts are composed of two main groups of tissues: i) the stroma, which consists of adipose tissue, connective tissue and extracellular matrix, and ii) the epithelium, which contains a network of ducts and lobules (Hassiotou & Geddes, 2013). Both the male and female breasts have glandular tissues that specialise in milk production; however in females, the glandular tissues are more developed (Hassiotou & Geddes, 2013). Each female breast contains 15 to 20 lobes, and within each lobe, there are 20 to 40 smaller sections called lobules or terminal duct lobular units (TDLUs) (Hassiotou & Geddes, 2013). TDLU is the functional unit of female breasts and is composed of milk-producing glands. During lactation, milk is produced in these lobules or TDLUs that travels through a network of tubes called ducts and exit the skin through the nipple. The space between lobes in the breast is mainly occupied by adipose tissues. Female breasts are supported by connective tissues and ligaments that also define their shape. A network of lymph vessels and lymph nodes is also present in the breast (Hassiotou & Geddes, 2013) (Figure 1.1).

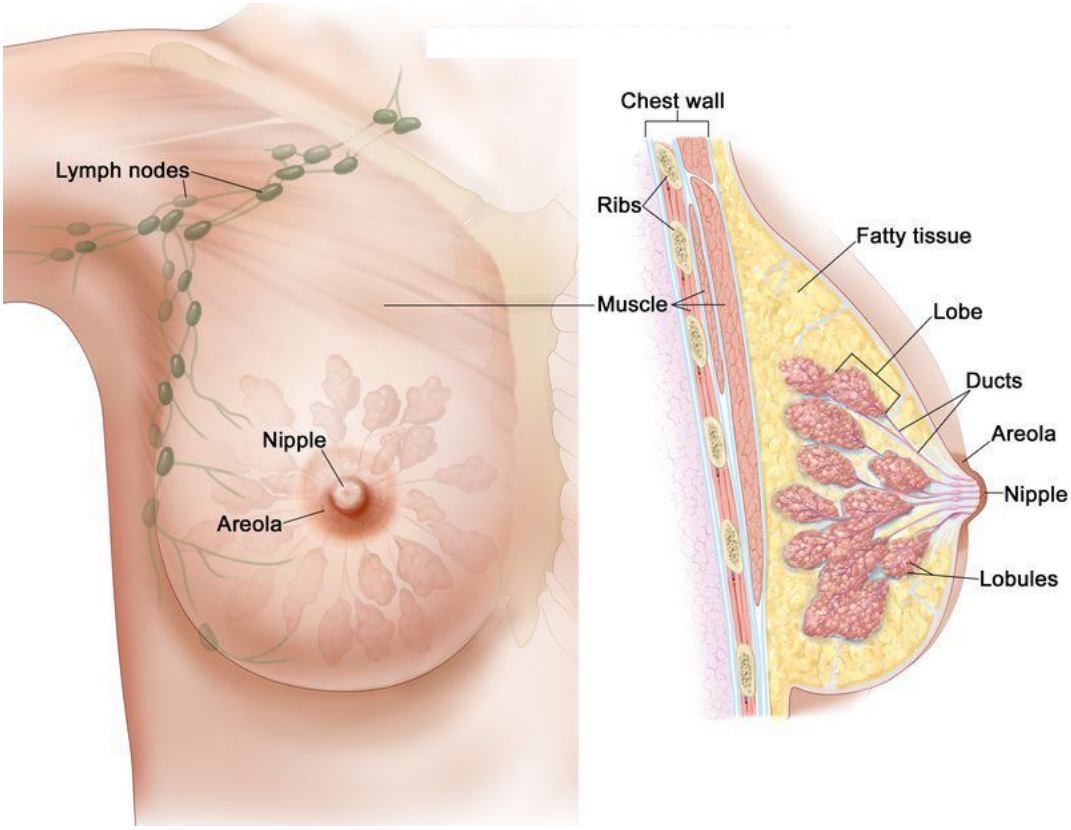


Figure 1.1: Anatomy of the human female breast from The National Breast Cancer Foundation (National Breast Cancer Foundation, 2020) © Terese Winslow LLC, U.S. Govt. has certain rights.

The breast is a tissue overlying the pectoral muscles. Female breasts are made of glandular tissues that produce milk and fatty tissues that determine the size of the breast. The milk-producing unit of the breast is organised in lobes (15-20) that further divide into lobules where milk is produced. The milk travels through a network of ducts and eventually exit through the nipple. The dark area of skin surrounding the nipple is called the areola. Connective tissue and ligaments provide support to the breast and define its shape. The breast also contains blood vessels, lymph vessels, and lymph nodes.

Female breasts undergo more changes than any other part of the human body. At puberty, female breasts develop in response to the female hormones estrogen and progesterone. The proliferation of epithelial and connective tissues and increased deposition of adipose tissues (fat) lead to breasts enlargement. At the time of pregnancy, female breasts further increase in volume and density in response to estrogen, progesterone, prolactin and placental hormones and develop secretory cells that produce milk proteins (Hassiotou & Geddes, 2013). After pregnancy, estrogen and progesterone levels decrease, while the levels of prolactin and growth hormones increase, leading to the production and secretion of milk. Once lactation ceases, female breasts decrease in size due to the degeneration of glandular tissues, ducts and lobules (Hassiotou & Geddes, 2013). With ageing and at menopause, ducts and lobules decrease in number and female breasts predominantly contain fat and stroma thus, leading to a reduction in the breast size (Hassiotou & Geddes, 2013).

1.2 The global trend in breast cancer incidence and mortality

Breast cancer is the most commonly diagnosed cancer and the leading cause of mortality in women worldwide. It is estimated that 1 in 7 women in Australia will be diagnosed with breast cancer in their lifetime (Cancer Council Australia, 2020). According to the recent data from GLOBOCAN 2018, there were more than 2 million newly diagnosed cases of breast cancer among women worldwide and an estimated 626,700 breast cancer deaths in over 100 countries in 2018 (Bray *et al.*, 2018).

Breast cancer incidence has constantly been rising over the past decades in most countries around the world. The global breast cancer incidence increased from 641,000 cases in 1980 to over 2 million cases in 2018, and this trend is likely to continue (Bray *et*

al., 2018). The highest incidence rates were recorded for developed regions such as Australia and New Zealand (age-standardised rate (ASR), 94.2 per 100,000), Western Europe (ASR, 92.6 per 100,000), Northern Europe (ASR, 90.1 per 100,000) and Northern America (ASR, 84.8 per 100,000) (Bray *et al.*, 2018). Increased awareness and a higher number of women undergoing regular breast screening in these regions, along with other risk factors such as lifestyle, dietary habits and environmental factors may be the likely explanation for an increased incidence. In contrast, many developing countries in Asia, Africa, and Latin America, where the incidence rates have been historically lower, are also witnessing a rapid increase in the number of breast cancer cases (Bray *et al.*, 2004). This trend may be related to changes in socio-economic conditions and changes in lifestyle and behaviour of women in these regions such as late pregnancy, having fewer children, reduced physical activity leading to obesity, increased awareness of breast cancer risks and regular breast screening.

Although the incidence rates have increased, the mortality rates of breast cancer have declined since 1993 in developed countries like the USA, Canada, Australia and many European countries primarily because of the availability of early detection techniques and advanced treatment options (Bray *et al.*, 2018). In contrast, the mortality rates remained high in many countries in Asia, Africa and Latin America (Bray *et al.*, 2018). Lack of breast cancer screening, late-stage diagnosis and lack of adequate treatment facilities are likely to contribute to this increase. The highest estimated mortality rates worldwide have been reported in Fiji, Melanesia (ASR, 25.5 per 100,000) (Bray *et al.*, 2018).

1.3 Risk factors of breast cancer

Breast cancer is a multifactorial disease, and several factors can increase a woman's chance of developing this disease such as age, ethnicity, lifestyle and family history. Some of the risk factors of breast cancer are detailed in the sections below.

1.3.1 Age

Breast cancer risk increases proportionally with increasing age with a sharp increase in risk observed after the age of 40 years (Bray *et al.*, 2018). In Australia, women aged 50 years are estimated to be at ten times increased risk of developing breast cancer than women aged 30 years (AIHW, 2014). The median age of breast cancer diagnosis varies across the world, with the peak age at breast cancer diagnosis estimated to be 40-50 years in Asian countries, 60-70 years in western countries and approximately 45 years in African countries (Bray *et al.*, 2018).

1.3.2 Reproductive history

Exposures to female hormones (both endogenous and exogenous) are known to influence breast cancer risk. Several factors that influence the hormonal exposures such as early onset of menarche, late menopause, delayed first birth and low parity are associated with risk of breast cancer (Kelsey *et al.*, 1993).

A large, pooled analysis of data from 117 international studies conducted between 1970 and 1999 including 118,964 women with invasive breast cancer and 306,091 controls, reported an increased risk of 1.05 (95% confidence interval (CI):1.04-1.06) for each year younger at menarche and an increased risk of 1.03 (95% CI: 1.02-1.03) for each year older at menopause. Both these associations were found to be stronger for invasive

lobular breast cancer (ILBC) compared with invasive ductal breast cancer (IDBC) (*P*-value < 0.006) (Collaborative Group on Hormonal Factors in Breast, 2012). Both early menarche and late menopause increase the lifetime exposure to estrogen, which may be part of the explanation for the increased risk of breast cancer.

Nulliparity (no history of giving birth) has been associated with an increased risk of breast cancer, whereas parity (history of giving birth) has been associated with a decreased risk of breast cancer. A pooled analysis of data from 47 epidemiological studies including 50,302 women with invasive breast cancer and 96,973 controls, reported that women with breast cancer had, on average, fewer childbirths than the controls (2.2 *versus* 2.6) (Collaborative Group on Hormonal Factors in Breast Cancer, 2002). Another pooled analysis of data from 17 studies; 3 cohorts, 13 case-control and one nested case-control study reported that women who have not had any children had an increased risk of breast cancer (relative risk (RR) = 1.16, 95% CI: 1.04-1.26) compared with women who have had at least one child (Nelson *et al.*, 2012). It was also noted that each childbirth reduced the risk by 7% especially for the estrogen receptor (ER) positive subtypes (Nelson *et al.*, 2012).

There is considerable evidence that women who are older at the birth of their first child are at an increased risk of breast cancer compared with women who are of younger ages. Data from a recent study including 3,768 women with breast cancer showed that a woman's risk of breast cancer increases by 3% for each year older she is at the first full-term pregnancy (RR = 1.03, 95% CI: 1.02-1.03) (Sisti *et al.*, 2016). The RR for breast cancer in women aged 30 years and older at first birth compared with women age 25-29 years at first birth has been estimated to be 1.20 (95% CI:1.02-1.42) (Nelson *et al.*, 2012).

Breastfeeding has been weakly associated with a decreased risk of breast cancer in the mother (Martin *et al.*, 2005; Wise *et al.*, 2009). The RR for breast cancer has been reported to be 0.61 (95% CI:0.44-0.85), in women who have breastfed compared with women who have never breastfed (Ying Zhou *et al.*, 2015). There is also evidence of a dose-response relationship, which means that women who have breastfed for a longer duration have proportionally larger reduced risk, with the risk of breast cancer estimated to be 0.98 (95% CI: 0.97-0.99) for per 5-months increase in the duration of breastfeeding (Chan *et al.*, 2019).

1.3.3 Menopausal hormone therapy

Use of combined estrogen and progesterone menopausal hormone therapy (MHT) increases a woman's risk of developing breast cancer. An increased risk of 1.33 (95% CI:1.30-1.36) has been reported for women who have used MHT compared with women who have never used MHT and the risk was estimated to be higher in current users compared with past users (RR = 1.72, 95% CI:1.55-1.92) (Munsell *et al.*, 2014). The duration of MHT use has also been reported to have an influence, with an increase in breast cancer risk with the duration of MHT use and higher among women who started using combined MHT close to menopause (Beral *et al.*, 2011).

1.3.4 Mammographic breast density

Mammographic breast density is assessed by mammography and can be expressed as percentage mammographic density (PMD). Variations in PMD are related to the difference in the amount of collagen and number of epithelial and non-epithelial cells in the breast (AT Wang *et al.*, 2014). It has been estimated that women with dense breasts (breasts with a higher proportion of epithelial and connective tissues than adipose tissues)

are 1.53 times more likely to develop breast cancer than women with average breast density (Pettersson *et al.*, 2014).

Breast density has a strong genetic component, and up to 60% of the variations in breast density can be explained by heritability (Boyd *et al.*, 2002). It is also associated with other risk factors of breast cancer. For instance, breast density was found to be inversely associated with parity (*P*-value = 0.02) and this association was stronger in women with a history of hormone therapy usage (*P*-value < 0.001) (Yaghjyan *et al.*, 2012).

1.3.5 Family history and genetic factors

A family history of breast cancer has been shown to influence a woman's risk of developing breast cancer (Pharoah *et al.*, 1997). The increased breast cancer risk associated with one affected first-degree relative has been estimated to be 1.80 (95% CI: 1.70-1.91), with two affected first-degree relatives to be 2.93 (95% CI: 2.37-3.63) and with three or more affected first-degree relatives to be 3.90 (95% CI: 2.03-7.49), compared with women with no family history of breast cancer (Collaborative Group on Hormonal Factors in Breast, 2001). The RR associated with having one or more affected second-degree relatives has been estimated to be 1.5 (95% CI:1.4-1.6) (Pharoah *et al.*, 1997). Similar RRs have been reported by other studies (Bevier *et al.*, 2012; Kharazmi *et al.*, 2014; Beebe-Dimmer *et al.*, 2015). It was also noted that the increased risk associated with having a first-degree relative is likely to be higher for younger women and for women, whose relative was diagnosed with breast cancer at a younger age (Collaborative Group on Hormonal Factors in Breast, 2001). Family history of other cancers such as ovarian cancer, prostate cancer and colorectal cancer are also known to increase a

women's risk of developing breast cancer (Beebe-Dimmer *et al.*, 2015). Breast cancer risk has been estimated to increase by 2.36 (95% CI: 1.59-3.37) for women with one or two first-degree relatives with ovarian cancer (Sutcliffe *et al.*, 2000).

Around 15-20% of breast cancers are familial (Pharoah *et al.*, 2002). The heritable component in familial breast cancers has been estimated to be 73%, whereas 27% has been considered to be due to the environmental factors (Couto & Hemminki, 2007). The heritable component is further higher in women with younger age at diagnosis (Couto & Hemminki, 2007). Inherited pathogenic variants in breast cancer susceptibility genes can partially explain the increased familial risk for breast cancer. Breast cancer susceptibility genes and the pathogenic variants in them are classified into further categories based on their minor allele frequency (MAF) and conferred risk as: i) high-risk variants (very rare, MAF < 0.005, conferred RR of breast cancer > 4), ii) moderate-risk variants (rare, MAF = 0.005-0.01, conferred RR = 2-4), and iii) low-risk variants (MAF > 0.01, conferred RR < 1.5) (Mavaddat *et al.*, 2010). High-risk variants in *BRCA1* (Miki *et al.*, 1994), *BRCA2* (Wooster *et al.*, 1995), *TP53* (Malkin *et al.*, 1990), *STK11* (A Hemminki *et al.*, 1998), *CDH1* (Berx *et al.*, 1995) and *PTEN* (J Li *et al.*, 1997) account for ~20% of the familial risk (Ghoussaini & Pharoah, 2009). Moderate-risk variants in *CHEK2* (Meijers-Heijboer *et al.*, 2002) and *ATM* (Renwick *et al.*, 2006) account for up to 5% of the familial risk. Variants in *PALB2* were initially classified as moderate-risk variants (Rahman *et al.*, 2007). However, more recent research suggests that it confers a similar risk as *BRCA2* (A Antoniou *et al.*, 2003; AC Antoniou *et al.*, 2014).

Genome-wide association studies (GWAS) have been a standard approach for identifying low-risk breast cancer susceptibility loci (Douglas F. Easton *et al.*, 2007). It

allows hundreds of thousands of common genetic variants or single nucleotide polymorphisms (SNPs), with a MAF of 1% or more to be analysed in case-control studies. To date, more than 170 breast cancer low-risk loci have been identified that explain ~18% of the familial risk of breast cancer (Douglas F Easton *et al.*, 2007; Thomas *et al.*, 2009; W Zheng *et al.*, 2009; Turnbull *et al.*, 2010; RL Milne *et al.*, 2014; Michailidou *et al.*, 2015; Michailidou *et al.*, 2017; Zhang *et al.*, 2020). Taken together, these currently known risk variants and loci only explain less than half of the genetic risk of familial breast cancer and the genetic component for the remaining half is still unknown (AC Antoniou & Easton, 2006).

BRCA1 and *BRCA2* are the most prominent of all the breast cancer susceptibility genes, accounting for most of families with multiple cases of breast and ovarian cancer (A Antoniou *et al.*, 2003). They are involved in several cellular pathways such as cell cycle, transcriptional regulation, apoptosis and DNA repair mechanism such as homologous recombination of double-stranded breaks (Roy *et al.*, 2012). The frequencies of *BRCA1* and *BRCA2* pathogenic variants in the general population have been estimated to be 1 in 400 and 1 in 800, respectively (Ford *et al.*, 1994; Claus *et al.*, 1996). A large case-control study, using a panel of 25 genes tested 95,561 women for hereditary breast cancer risk and estimated the increased risk of breast cancer associated with *BRCA1* mutations as 5.91 (95% CI: 5.25-6.67) and with *BRCA2* mutations as 3.31 (95% CI: 2.95-3.71) (Kurian *et al.*, 2017). Kuchenbaecker et al., (2017) in a prospective study, reported that for *BRCA1* mutation carriers, the incidence increased rapidly with age, up to the age of 41-50 years and remained constant after that, throughout the remaining lifetime, whereas for *BRCA2* mutation carriers, the incidence kept increasing up to the age of 51-60 years (5-10 years later than *BRCA1* mutation carrier) (Kuchenbaecker *et al.*, 2017).

The cumulative risk of breast cancer to age 80 years, for *BRCA1* and *BRCA2* mutation carriers, has been estimated to be 72% (95% CI: 65%-79%) and 69% (95% CI: 61%-77%), respectively (Kuchenbaecker *et al.*, 2017).

1.4 Classification of breast cancer

The heterogeneity of breast cancer is evident in its various molecular features and therapeutic responses. This heterogeneity poses a pronounced challenge in the management of the disease. Classifying breast cancer into clinically relevant subtypes has the potential for more accurate prognostication and precision medicine by identifying subtype-specific diagnostic and prognostic markers and therapeutic targets. Traditional methods for breast cancer classification are based on the clinicopathological features of the tumour, mainly tumour grade and tumour stage (Rakha, Reis-Filho, Baehner*, et al.*, 2010). Although these methods are well validated and still have clinical value, they fail to cover the broad and diverse spectrum of breast cancer heterogeneity. With advancements in molecular technologies, breast cancer classification has evolved from traditional histopathological classification to a more advanced sub-classification system based on the molecular level information (Rakha & Green, 2017). Different approaches to breast cancer classification are detailed below.

1.4.1 Histology, grade and stage

Histology refers to the growth pattern of tumours and numerous breast cancer subtypes have been defined based on the tumour morphological characteristics. The most common histological subtype of breast cancer is invasive ductal breast cancer (IDBC), also known as Invasive Carcinoma of No Special Type. It accounts for 40-70% of all breast cancer cases and is a heterogeneous group of diseases that do not show sufficient

special differentiation pattern to be classified as the special subtype (Lakhani *et al.*, 2012). The remaining 25-30% of breast cancers are classified as special subtypes based on their specific growth patterns (Lakhani *et al.*, 2012). According to the 4th edition of the WHO classification of breast cancers published in 2012, more than 18 special subtypes of breast cancer have been described (Lakhani *et al.*, 2012). Of those, invasive lobular breast cancer (ILBC) is the most common special subtype accounting for 10-15% of all cases. Other special breast cancer subtypes include tubular carcinoma, cribriform carcinoma, mucinous carcinoma, neuroendocrine carcinoma, micropapillary carcinoma, papillary carcinoma, carcinoma with medullary properties, and metaplastic carcinoma (Lakhani *et al.*, 2012). Although the breast cancer subtypes defined based on histology have been shown to display different biological and clinical behaviour (Weigelt *et al.*, 2008), the implication of histological classification has limited precision due to the subjective nature of the histopathological imaging. The heterogeneity of breast cancers in a single subtype further limits the clinical utility of the histological classification.

Grade represents the degree of differentiation of tumour tissues and is one of the most important prognostic factors that inform about the tumour clinical behaviour (Rakha, Reis-Filho, Baehner, *et al.*, 2010). Nottingham grading system is a widely accepted system that assigns tumour grades by evaluating the morphological features such as degree of tubule formation, degree of nuclear pleomorphism and tumour mitotic count (Elston & Ellis, 1991). Tumours are scored based on the above-mentioned parameters and graded as grade I (total score between 3 to 5), grade II (total score 6 or 7) and grade III (total score 8 or 9) (Elston & Ellis, 1991). Tumour grade has been reported to be an independent prognostic factor (Saimura *et al.*, 1999; Rakha *et al.*, 2008; Wirapati *et al.*, 2008), used in many clinical prediction tools such as the Nottingham prognostic

index and Adjuvant online, together with other biomarkers to predict patient prognosis in breast cancer (Olivotto *et al.*, 2005; Fong *et al.*, 2015). High-grade breast tumours are associated with higher rates of metastasis, higher rates of recurrence and poor patient outcome compared with low-grade breast tumours (Rakha, Reis-Filho, Baehner*, et al.*, 2010).

Breast tumour staging is another factor, which is used to inform about the tumour status, patient prognosis and to guide treatment decisions. The TNM system for breast cancer staging is a globally accepted system established by The American Joint Committee on Cancer (AJCC) in 1977. It assigns the tumour stage based on the anatomic extent of the disease, including tumour size (T), nodal involvement (N) and distant metastasis (M) (Edge *et al.*, 2010). Although the anatomical staging remains an important factor in clinical decision making, it is purely based on the anatomical features of the tumour and does not account for other important tumour features that have significant prognostic and predictive values such as tumour grade and biological factors such as estrogen receptor (ER), progesterone receptor (PR) and human epidermal growth factor receptor 2 (HER2) expression status (see section 1.4.2). Therefore, the recent edition of the AJCC has established a new prognostic staging system that incorporates other biological features and also incorporates molecular level information (Hortobagyi *et al.*, 2017). Savage et al., (2019) in a retrospective cohort study including 1,703 women, showed that the prognostic staging system displayed improved prognostic accuracy with respect to the overall survival compared with the anatomical staging system (Savage *et al.*, 2019). A similar finding has been reported by other studies (SB Lee *et al.*, 2018; Weiss *et al.*, 2018).

1.4.2 Hormone receptor and HER2 expression status

In routine clinical practice, breast cancer is commonly classified based on the expression status of ER, PR and HER2 (Rakha, Reis-Filho, & Ellis, 2010), which is commonly determined by immunohistochemistry (IHC) or fluorescent in-situ hybridisation and the tumours are termed positive or negative based on the expression status.

ER expression status has been used for the clinical management of breast cancer since the 1970s, where it was originally used to predict endocrine therapy response and as a prognostic biomarker for early recurrence (Rakha, Reis-Filho, & Ellis, 2010). Most breast tumours (~80%) are estimated to be ER positive (Anderson *et al.*, 2002). ER positive breast tumours are typically well-differentiated, display less aggressive tumour features and are associated with better short-term patient outcome compared with ER negative breast tumours (Dunnwald *et al.*, 2007). A difference in gene expression profile has been reported based on the ER positive or ER negative status of breast tumours (Perou *et al.*, 2000; Sørlie *et al.*, 2001).

Around 55-65% of breast tumours are estimated to be PR positive and are associated with a better prognosis compared with PR negative breast tumours (Rakha, Reis-Filho, & Ellis, 2010). PR expression status is generally used in combination with ER expression status for classifying breast tumours and based on this combinatorial approach four subtypes of breast tumours have been defined; i.e., ER positive/PR positive, ER positive/PR negative, ER negative/PR positive and ER negative/PR negative. Among these, the ER positive/PR positive subtype represents 55-65% of breast tumours. The women with these tumours are usually of older age, present with lower grade and smaller size tumours and have a better prognosis than women with other tumour

subtypes (Dunnwald *et al.*, 2007; Rakha *et al.*, 2007). Around 75-85% of the ER positive/PR positive breast tumours show a good response to hormone therapy (Dowsett *et al.*, 2006). In contrast, the ER negative/PR negative subtype accounts for 20-25% of the breast tumours and are associated with higher grade, higher recurrence rates, worse prognosis and poor response to hormone therapy (Kinne *et al.*, 1987; Anderson *et al.*, 2001; Bardou *et al.*, 2003). Breast tumours, which are positive for a single hormone receptor; i.e., from the ER negative/PR positive or ER positive/PR negative subtypes, are generally characterised by higher grade and larger size tumours, respond poorly to endocrine therapy and show enhanced expression of EGFR and HER2 compared with the ER positive/PR positive subtype (Bardou *et al.*, 2003; Rakha *et al.*, 2007). The ER positive/PR negative subtype accounts for ~17% of breast cancer cases, whereas the ER negative/PR positive tumours are relatively rare, accounting for 0.2-10% of all cases (Anderson *et al.*, 2001; Dunnwald *et al.*, 2007; Rakha *et al.*, 2007).

HER2 gene amplification and protein overexpression has been shown to have predictive value for response to targeted anti-HER2 therapy (Bartlett *et al.*, 2003; Kaufmann *et al.*, 2006; Wolff *et al.*, 2007). Around 13-20% of breast tumours are HER2 positive, and of these, 50% are ER and PR negative (Slamon *et al.*, 1987; Dandachi *et al.*, 2002). The ER negative/PR negative and HER2 positive breast tumours show a poorer prognosis compared with the ER positive/PR positive and HER2 negative subtype (Rakha, Reis-Filho, & Ellis, 2010). Around 10-15% of breast tumours are ER negative/PR negative and HER2 negative (Vuong *et al.*, 2014). This subtype is often referred to as the triple negative breast cancer (TNBC). They represent a distinct group of tumours with completely different clinical presentation and patient prognosis. Women with TNBC have typically been reported to be of younger age and present with higher grade and larger size

tumours compared with women with other breast cancer subtypes (Dent *et al.*, 2007). TNBC tumours were also found to be associated with an increased risk of distant recurrence (hazard ratio (HR): 2.6, 95% CI: 2.0-3.5, *P*-value < 0.001) and a poorer 5-years survival (HR: 3.2, 95% CI: 2.3-4.5, *P*-value < 0.001) (Dent *et al.*, 2007). The spectrum of TNBC is diverse, and it exemplifies the heterogeneity present in breast cancer subtypes.

In addition to ER, PR and HER2, other biomarkers have been proposed to be used in breast cancer classification including ER-ß, androgen receptor, proliferation-related marker Ki-67, cytokeratin5/6 and EGFR (Vuong *et al.*, 2014). Ki-67 has been shown to further refine ER positive breast tumour group, and a combined group of biomarkers (ER, PR, HER2 and Ki-67) has been proved to be useful in clinical settings (Cuzick *et al.*, 2011). These biomarkers are being used more frequently to further refine the current breast cancer classification (Blows *et al.*, 2010; Green *et al.*, 2013).

1.4.3 Gene expression profiling-based classification

The development of gene expression arrays enabled a more detailed analysis of genes expressed by breast tumours compared with IHC. In 2000, Perou et al. performed complementary DNA microarray gene expression analysis of 8,102 genes in 38 invasive breast cancer cases that included 36 IDBC and two ILBC, one ductal carcinoma in-situ (DCIS), one fibroadenoma and four normal breast samples, and defined four main intrinsic subtypes of breast cancer: i) luminal ii) basal-like iii) HER2-enriched and iv) normal like (Perou *et al.*, 2000). A subsequent study from the same group further sub-classified the luminal subtype into luminal A and luminal B differing in the expression of proliferation related genes and also demonstrated that these subtypes were different in terms of their clinical outcome (Sørlie *et al.*, 2001).

Luminal A tumours tend to express ER and PR and are more likely to be negative for HER2. Of the four intrinsic subtypes, luminal A subtype has been shown to have the best prognosis with relatively lower recurrence and higher survival rates (Carey *et al.*, 2006; Dawood *et al.*, 2011). In contrast, luminal B tumours are ER positive and/or PR positive and are highly proliferative with a high Ki-67 score. The proliferative nature of luminal B tumours has been suggested to be the likely cause of poorer prognosis in comparison with luminal A tumours (Calza *et al.*, 2006; Carey *et al.*, 2006). Luminal B tumours are typically of higher grade, larger size and are more likely have *TP53* mutations compared with luminal A breast tumours (Calza *et al.*, 2006; Carey *et al.*, 2006; Dawood *et al.*, 2011).

The basal-like tumours represent the most heterogeneous subtype with respect to the histopathological features, mutation profiles, clinical behaviour and patient outcome (Sørlie *et al.*, 2001; Banerjee *et al.*, 2006; Chin *et al.*, 2006; Fulford *et al.*, 2007). They are commonly characterised as grade III tumours with a high proliferative index, aggressive clinical course and frequent somatic pathogenic variants in *TP53* (Badve *et al.*, 2011). Basal-like breast cancers are a subset of TNBC in which the tumours also express high molecular weight cytokeratin such as cytokeratin5/6 and cytokeratin14 that are usually expressed in normal breasts (Rakha, Reis-Filho, & Ellis, 2010).

The HER2 enriched tumours are typically negative for ER and PR expressions and overexpress HER2. They are associated with high tumour grade, early and frequent metastases and poor prognosis (Calza *et al.*, 2006; Carey *et al.*, 2006). HER2 enriched tumours are known to benefit from antibody-based and anti-kinase based therapies that target HER2 and has changed the prognosis of this subtype (Slamon *et al.*, 2001; Piccart-Gebhart *et al.*, 2005).

The normal-like subtype is characterised by the expression of genes associated with adipose tissues and other stromal cell types. This subtype is thought to represent normal cell contamination of samples and remain a controversial group (Vuong *et al.*, 2014).

There are further breast cancer subtypes identified since the first gene-expression based classification. These are claudin-low, molecular apocrine and interferon-related subtypes. The claudin-low subtype is commonly enriched for metaplastic carcinoma and tumours with medullary-like features. The tumours in this subtype are usually triple negative and show enrichment for epithelial-to-mesenchymal transition markers, immune response genes and cancer stem cell-like features (Prat *et al.*, 2010). The survival rates for claudin-low breast cancer subtype lie between those for luminal and basal-like subtype (Prat *et al.*, 2010). The molecular apocrine subtype is characterised by ER positive and androgen receptor positive tumours with characteristic histological features related to apocrine breast cancer subtype (Farmer *et al.*, 2005). They are associated with early recurrence; however show a good response to neoadjuvant chemotherapy (Guedj *et al.*, 2012). The interferon-related subtype is characterised by high expression of interferon-regulated genes, including *STAT1* (Z Hu *et al.*, 2006).

Although gene expression-based stratification of breast cancer expanded our knowledge of breast cancer heterogeneity, its clinical application in identifying high-risk patients and tailoring therapy remains limited. Several prognostic signatures have been developed based on the expression profiling however, they have not been able to completely replace the traditional pathological classification system. One of the main reasons for this is the lack of reproducibility and the absence of a standard methodology for these classifications. Furthermore, most of the gene expression-based studies relate to IDBC as they constitute the majority of breast cancer cases. Data relating to the

expression profiles of special subtypes of breast cancer are limited. Molecular classification of the breast tumour is still evolving and there is a growing interest to incorporate other molecular level information for a more comprehensive view of the breast cancer heterogeneity.

1.5 Invasive Lobular Breast cancer

Lobular breast cancer was first reported by Foote and Stewart in 1941, where they defined both in-situ and invasive forms of the disease (Foote & Stewart, 1941). ILBC is the second most common histological subtype of breast cancer accounting for 10-15% of all cases (Lakhani *et al.*, 2012). After its first report, ILBC remained largely underrepresented in major breast cancer studies mainly because of its low prevalence compared with the more common histological subtype, IDBC. Recent research focus has shifted to ILBC, and it is increasingly recognised as a distinct breast cancer subtype with different tumour biology and clinical presentation (Arpino *et al.*, 2004; C. I. Li *et al.*, 2005). Further investigation of its unique tumour biology and molecular features has been pursued to identify diagnostic and prognostic biomarkers specific for this subtype.

1.5.1 Epidemiology and risk factors of ILBC

ILBC is more likely to occur in older women with the median age of diagnosis being 65 years (Arpino *et al.*, 2004). The incidence rate of ILBC increased sharply (1.52-fold, 95% CI: 1.42-1.63) between 1987 to 1999, which was mainly related to increased use of hormone replacement therapy (HRT) in the USA during that time (E Hemminki *et al.*, 1988; Christopher I. Li *et al.*, 2003). Awareness of the increased risk of breast cancer associated with HRT led to its reduced use and a decline in ILBC incidence rate (Christopher I. Li & Daling, 2007). The annual percent change in ILBC incidence rate

fell steadily from 1998 to 2004 (Annual percent change = -3.18%, 95% CI: -5.18 to -1.03) (Christopher I. Li & Daling, 2007). However, recent data suggests an increase in its incidence rate again (Bray *et al.*, 2018).

Several breast cancer risk factors particularly those related to female hormone exposures (both endogenous and exogenous) are more strongly associated with ILBC than other breast cancer types (Collaborative Group on Hormonal Factors in Breast, 2012). Kotsopoulos et al., (2010) investigated the association between hormonal exposures and breast cancer risk in 4,655 IDBC and 659 ILBC tumours and reported a significantly stronger association between age at menarche (*P*-value = 0.03), age at first birth (*P*-value < 0.001) and postmenopausal hormone use (*P*-value < 0.001) in ILBC compared with IDBC (Kotsopoulos *et al.*, 2010). Breast cancer risk associated with older age at first birth (>=30 years compared with < 20 years) was found to be more pronounced for ILBC (odds ratio: 2.4, 95% CI: 1.90-2.90) compared with IDBC (odds ratio: 1.3, 95% CI: 1.20-1.40), *P*-value = 0.01 (Newcomb *et al.*, 2011). Increased risk associated with nulliparity compared with women with age at first birth < 20 years was also found to be stronger for ILBC (odds ratio: 1.7, 95% CI: 1.34-2.20) compared with IDBC (OD: 1.2, 95% CI: 1.1-1.3), *P*-value = 0.004 (Newcomb *et al.*, 2011). Another large meta-analysis based on 14,102 breast cancer cases including 1,526 ILBC cases, observed that the RRs of breast cancer in current HRT users compared with women who have never used HRT was larger for ILBC (RR = 2.3, 95% CI: 2.0-2.5) and tubular breast cancer (RR = 2.7, 95% CI: 2.2-3.3) compared with mixed ductal lobular (RR = 2.1, 95% CI: 1.7-2.7), IDBC (RR= 1.6, 95% CI: 1.5-1.7) and mucinous breast cancers (RR = 1.6, 95% CI: 1.1-2.3). The RRs varied significantly according to tumour histology overall (P-value < 0.0001) (Reeves *et al.*, 2006).

ILBC has been shown to have higher familial risk compared with IDBC. Allen-Brady K *et al*., (2005) reported that the relatives of women affected by ILBC had an increased risk for ILBC (first-degree relative: familial relative risk (FRR) = 4.51, 95% CI: 2.8-6.9) and an increased risk for any type of breast cancer (first-degree relative: FRR = 2.5, 95% CI: 2.1-2.9) (Allen-Brady *et al.*, 2005). These estimates were significantly higher than all breast cancer and early-onset breast cancer FFR estimates (1.83, 95% CI: 1.75-1.90 and 2.42, 95% CI: 2.21-2.63, respectively), suggesting a strong heritable component associated with this subtype (Allen-Brady *et al.*, 2005). Despite this, the only recognised breast cancer susceptibility gene linked to ILBC predisposition is *CDH1*. Germline pathogenic variants in *CDH1* was originally linked to the predisposition of hereditary diffuse gastric cancer (Guilford *et al.*, 1998). It was noted that many of these hereditary diffuse gastric cancer families with germline pathogenic variants in *CDH1* often had cases of ILBC (Keller *et al.*, 1999; Brooks-Wilson *et al.*, 2004; Suriano *et al.*, 2005) that led to the identification of *CDH1* as a susceptibility gene for ILBC as well. However, germline *CDH1* pathogenic variants predisposing to ILBC without a family history of HDGC are not common, reported in 1-4% of cases (Masciari *et al.*, 2007; Schrader *et al.*, 2011; Xie *et al.*, 2011).

Women with ILBC have been reported to be more commonly *BRCA2* mutation carriers (8-10% of cases), compared with *BRCA1* mutation carriers (up to 2% of cases) (Mavaddat *et al.*, 2012; Ditchi *et al.*, 2019). Low-risk loci predisposing to ILBC has also been identified. A GWAS including pooled data from 36 studies that included 5,622 ILBC, 401 lobular carcinoma in-situ (LCIS) and 34,271 unaffected women, identified a SNP at 7q34 specific to ILBC predisposition (Sawyer *et al.*, 2014).

1.5.2 Histological subtypes of ILBC

ILBC is highly varied in terms of its morphology, and several subtypes of ILBC have been described based on the growth pattern (Martinez & Azzopardi, 1979; Lakhani *et al.*, 2012). The most common form or subtype of ILBC is classic ILBC representing ~50% of all cases (Lakhani *et al.*, 2012). It is characterised by small, round and discohesive cells that grow in a single file without forming any distinct clusters. Classic ILBC cells maintain a uniform size, have pale cytoplasm with uniform and clear nuclei without any hyperchromatism. Mitotic structures in the nucleus are rare and nuclear pleomorphism (variation in nuclear size, shape, hyperchromatism, nucleoli) is low to moderate (Rakha & Ellis, 2010).

Tumours that lack the characteristic discohesive non-linear growth pattern of classic ILBCs, while still maintaining similar cytological features, are referred to as the histological subtypes of ILBC (Rakha & Ellis, 2010). Several subtypes have been identified such as alveolar, solid, pleomorphic, tubulolobular, signet ring cell, trabecular and mixed ILBC subtype. Figure 1.2 illustrates ILBC histological subtypes.

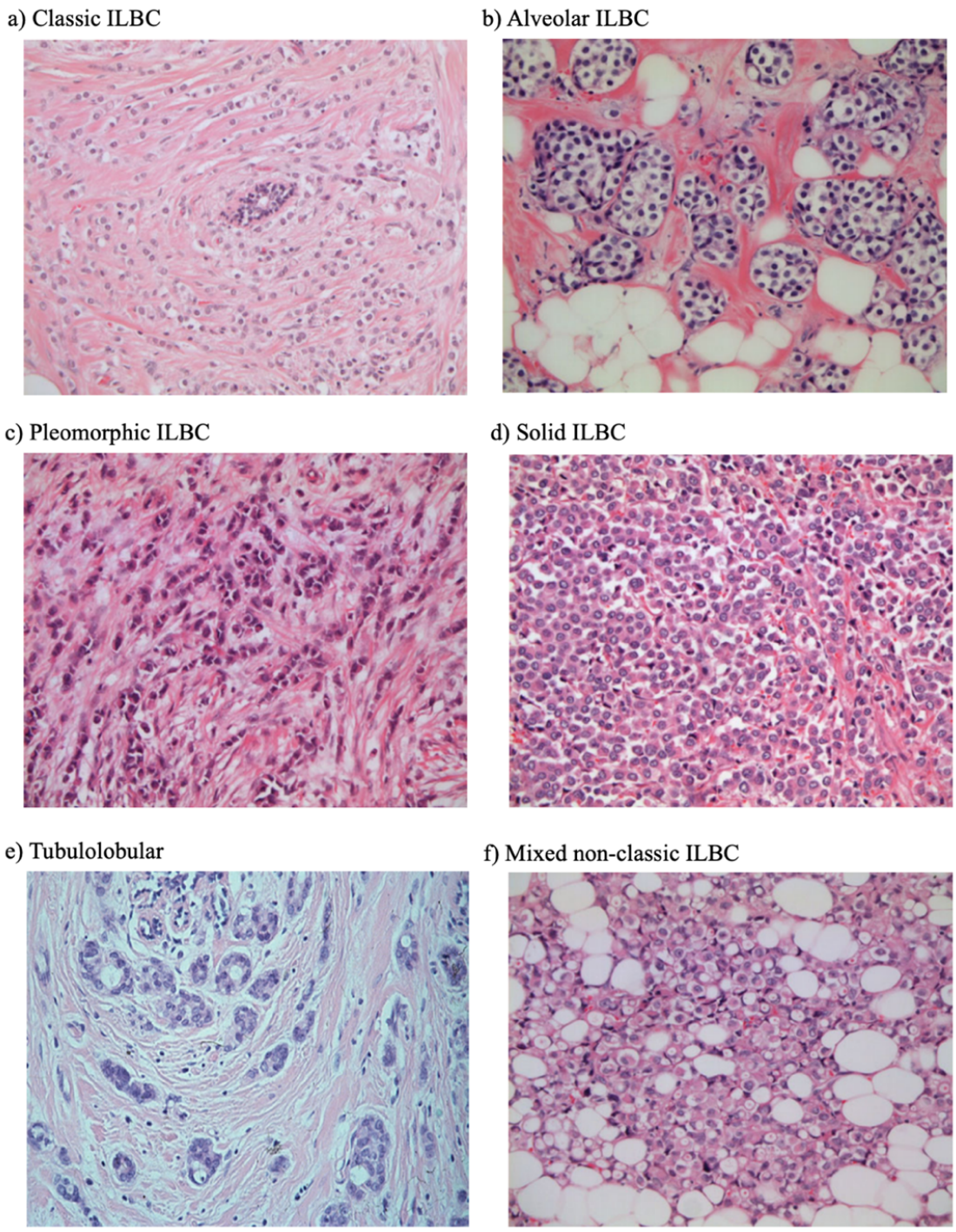


Figure 1.2: Histological subtypes of ILBC taken from (Iorfida *et al.*, 2012).

Representative images of hematoxylin and eosin-stained histological subtypes of ILBC, **a**) Classic-ILBC; **b**) Alveolar ILBC; **c**) Pleomorphic ILBC; **d**) Solid ILBC; **e**) Tubulolobular ILBC and **f**) ILBC mixed non-classic type.

The alveolar ILBC is characterised by an aggregation of cells (at least 20) that are arranged in glandular structures while the other cytological features of alveolar ILBC remains typical of classic ILBC (Shousha *et al.*, 1986). The solid ILBC subtype is characterised by typical small and non-cohesive cells of lobular morphology, but instead of the single file growth pattern of classic ILBC, these cells grow in solid sheets, are more pleomorphic (variable in shape and size) and have a higher frequency of mitosis compared with the classic subtype (Fechner, 1975). Pleomorphic ILBC is another subtype that shows a higher mitotic rate, cellular atypia and nuclear pleomorphism (nucleus with varying shape and size) similar to the solid subtype, while still retaining the characteristic growth pattern of the classic ILBC (Weidner & Semple, 1992; Middleton *et al.*, 2000). Pleomorphic ILBC accounts for less than 1% of all breast cancer cases and ~5% of all ILBC cases. They show more aggressive clinical behaviour than other ILBC subtypes with locally advanced tumours overexpressing HER2 and p53 (Middleton *et al.*, 2000). Pleomorphic ILBC is more frequently associated with LCIS and may show apocrine or histiocytoid differentiation (Eusebi *et al.*, 1992). Haque et al., (2019) comparing the clinical characteristics of classic ILBC (n = 114,859) and pleomorphic ILBC (n = 401) reported that pleomorphic ILBC tumours were more likely ER negative and HER2 positive and were associated with poorer overall survival compared with classic ILBC tumours (P-value < 0.001) (Haque *et al.*, 2019). The tubulolobular ILBC subtype has histological and molecular features common to both lobular and tubular carcinoma (a subtype of IDBC). It is usually characterised by low-grade tumours and associated with a good prognosis (Fisher *et al.*, 1977). The mixed ILBC subgroup represents the cases that present as a mixture of classic ILBC and one or more histological subtype of ILBC (J. M. Dixon *et al.*, 1982). The classic ILBC and the mixed ILBC subtype together

contribute to the majority of lobular tumours accounting for up to 75% of all ILBC cases (Orvieto *et al.*, 2008; Rakha *et al.*, 2008).

Apart from being different in terms of morphology, these ILBC subtypes have also been associated with specific genomic alterations. For instance, somatic variants in *ARID1A* and *TP53* have been more commonly reported in solid ILBC, whereas somatic variants in *HER2* and *TP53* have been commonly observed in pleomorphic ILBC (12-19% of pleomorphic ILBC) (Ercan *et al.*, 2012; Lien *et al.*, 2015; Deniziaut *et al.*, 2016; Zhu *et al.*, 2018; Rosa-Rosa *et al.*, 2019). *PYGM*, encoding the myophosphorylase enzyme, has also been found to be frequently mutated in pleomorphic ILBCs (Ciriello *et al.*, 2015).

1.5.3 Clinical features of ILBC

ILBC is generally associated with older age at diagnosis, higher tumour stage, lower histological grade and a higher percentage of multicentric, bilateral and multifocal tumours compared with IDBC (Arpino *et al.*, 2004; Chen *et al.*, 2017). ILBC tumours are typically ER and PR positive and HER2 negative (Arpino *et al.*, 2004; Chen *et al.*, 2017). HER2 overexpression has been reported in up to 5% of ILBC tumours (Desmedt *et al.*, 2016; Chen *et al.*, 2017). HER2 positive ILBCs are more likely to be negative for hormone receptors (ER and PR), have higher grade tumours and have poorer survival compared with HER2 negative ILBC tumours (Kee *et al.*, 2020). HER2 positive ILBC commonly belong to the pleomorphic ILBC subtype (Hoff *et al.*, 2002; Lal *et al.*, 2005).

The diffuse growth pattern of ILBC poses a major challenge for its early detection, which may be the likely explanation for women with ILBC more frequently presenting

with advanced-stage tumours (Arpino *et al.*, 2004; Molland *et al.*, 2004). ILBCs do not usually present with a firm lump in the breast and early signs may include thickening in parts of the breast, with or without any changes in the nipple (Biglia *et al.*, 2013). Compared with IDBC, mammography has a lower sensitivity for detecting ILBC ranging between 57-81% with a high false-negative rate of up to 19% (Hilleren *et al.*, 1991; Le Gal *et al.*, 1992; Krecke & Gisvold, 1993). In a recent study, Porter et al., (2014), reviewed 361 cases of ILBC diagnosed between 1995-2010 and found that the mammography was negative for 30% of the cases and the most common positive finding was a localised spiculated mass (Porter *et al.*, 2014). Figure 1.3 shows the mammography of a 70-year-old woman where ILBC can be visualised as a spiculated mass as reported by Porter et al., (2014).

Ultrasound is relatively more sensitive (78-95%) in detecting ILBC compared with mammography (Butler *et al.*, 1999; Chapellier *et al.*, 2000; Evans & Lyons, 2000; Cawson *et al.*, 2001), however there is a report suggesting an underestimation of tumour size on ultrasound that can have implications while surgery (Watermann *et al.*, 2005). Magnetic resonance imaging has been shown to have enhanced sensitivity for detecting ILBC, particularly the multifocal tumours however, the applicability of this technique to inform about breast conservation surgery in ILBC women is low (Schelfout *et al.*, 2004; Brem *et al.*, 2009; Barker *et al.*, 2019).

Figure 1.3: Mammography of a 70-year-old woman with ILBC from (Lopez & Bassett, 2009).

The mammography scan shows the left mediolateral mammogram of a 70-years-old woman. ILBC tumour can be seen as a low-density, irregular, spiculated mass (indicated by the arrow). The ILBC tumour in the mammogram presents with ill-defined margins and have equal opacity to the surrounding breast parenchyma.

ILBC tumours have a distinct pattern of metastasis. Mathew et al., (2017), comparing the metastatic patterns of ILBC and IDBC, found that compared with IDBC, ILBC showed a more frequent metastasis to bone (57% of 88 ILBCs *versus* 38% of 673 IDBCs, *P*-value = 0.001) and gastrointestinal tract (6% *versus* 0.3%, *P*-value < 0.001), and less frequent metastases to the lungs (6% *versus* 24%, *P*-value < 0.001) and the liver (5% *versus* 11%, *P*-value = 0.049), as the first site of distant recurrence (Mathew *et al.*, 2017). Over the entire course of the metastatic disease, more women with ILBC showed a metastatic preference for ovarian (6% *versus* 2%, *P*-value = 0.042) and gastrointestinal tracts (8% *versus* 0.6%, *P*-value < 0.001), whereas reduced tendencies to metastasise to the liver (21% *versus* 49%, *P*-value < 0.001) and lungs (24% *versus* 52%, *P*-value < 0.001) were observed (Mathew *et al.*, 2017). Another study, using The Surveillance Epidemiology and End Results 1990-2013 database, reported a higher frequency of bone metastasis in women with ILBC (92% of 85,048 ILBCs *versus* 76% of 711,287 IDBCs) (Chen *et al.*, 2017). ILBCs were also shown to have multiple metastatic sites (Chen *et al.*, 2017).

Current treatment protocol for ILBC is similar to other histological subtypes of breast cancer (Heilat *et al.*, 2019). In most cases of early breast cancers, breast-conserving surgery is performed to remove the tumour with a margin of the surrounding normal tissue (Heilat *et al.*, 2019). However, due to the diffuse and multifocal morphology of ILBC tumours, a higher mastectomy rate has been reported in women with ILBC (72% of 563 ILBCs *versus* 56% of 5,889 IDBCs, *P*-value < 0.01) (Pestalozzi *et al.*, 2008). Studies have reported that 17-65% of women undergoing breast-conserving surgery have been reported to undergo another surgery due to inaccurate estimation of the tumour margin (Silberfein *et al.*, 2010; Kryh *et al.*, 2014; Braunstein *et al.*, 2015; S Sharma *et al.*, 2015).

Women with ILBC are also more likely to undergo contralateral mastectomy because of the higher frequency of bilateral tumours in ILBC (Babiera *et al.*, 1997).

Neo-adjuvant chemotherapy has been consensually reported to be less effective in ILBC compared with IDBC (Cristofanilli *et al.*, 2005; Tubiana-Hulin *et al.*, 2006; Joh *et al.*, 2012; Delpech *et al.*, 2013; Loibl *et al.*, 2014). Loibl et. al., (2014) reported a significantly lower pathologic complete response rate (absence of invasive or in situ cancer in the breast and/or axillary lymph nodes) in women with ILBC compared with non-ILBC after neo-adjuvant chemotherapy (6% of 1,055 ILBCs *versus* 17% of 7,965 non-ILBCs, *P*-value < 0.001) (Loibl *et al.*, 2014). Another study reported a pathologic complete response rate of 1% of 118 ILBC cases compared with 9% of 742 IDBCs (*P*-value = 0.002) after neo-adjuvant chemotherapy (Tubiana-Hulin *et al.*, 2006). Neo-adjuvant chemotherapy has been advised for women with hormone receptor negative ILBC with advanced tumour stage (Loibl *et al.*, 2014).

As ILBC tumours typically express ER and PR, women with ILBC are considered to benefit from neoadjuvant endocrine therapy (Locker *et al.*, 1991; Tubiana-Hulin *et al.*, 2006). A retrospective study including 61 women with ER positive ILBC reported that neoadjuvant treatment with letrozole (an aromatase inhibitor) reduced the tumour volume by 66% (mean) in three months and 81% of the women had successful breast-conserving surgery after the neoadjuvant treatment (J Michael Dixon *et al.*, 2011).

There is a large variability in the literature concerning ILBC patient outcome compared with other breast cancer types. Women with ILBC have been reported to have a similar (Fortunato *et al.*, 2012), better (Cristofanilli *et al.*, 2005; Wasif *et al.*, 2010) or

no significant difference in prognosis compared with women with IDBC (Mhuircheartaigh *et al.*, 2008). Chen et al., (2017) comparing ILBC and IDBC diagnosed between 1990 and 2013, found that women with ILBC had an early advantage, showing better overall survival for up to 60 months (ILBC *versus* IDBC, HR= 1.12, *P*-value <0.0001), however long-term (after five years) prognosis for women with ILBC was poorer (ILBC *versus* IDBC, HR= 0.78; *P*-value <0.0001). Regarding the disease-free survival, women with IDBC showed better prognosis, both early and long-term, compared with women with ILBC (ILBC *versus* IDBC, HR = 0.81; *P*-value <0.0001) (Chen *et al.*, 2017).

1.5.4 The somatic genomic landscape of ILBC

Massively parallel sequencing technologies have enabled an in-depth molecular characterisation of breast cancer and have enhanced our understanding of the key genetic alterations in breast cancer tumourigenesis. However, most of the studies investigating the mutational landscape of breast cancer have focused mainly on IDBC (Cancer Genome Atlas, 2012; Curtis *et al.*, 2012).

In the first breast cancer study by The Cancer Genome Atlas (TCGA) where 466 breast tumours were assessed by whole-exome sequencing, only 8% (36/466) of all breast cancer cases were ILBC (Cancer Genome Atlas, 2012). Additionally, the analysis was mainly focused within the context of the gene expression-based breast cancer subtypes and the histological subtype-specific differences were not investigated (Cancer Genome Atlas, 2012). Somatic variants in *CDH1* and *HER2* were the only genetic alterations noted in ILBC tumours in this study (Cancer Genome Atlas, 2012). Another recent study profiled 2,433 primary breast tumours using targeted sequencing of 173 genes (Pereira *et*

al., 2016). ILBC represented 8% (194/2,433) of the total sample size and again inactivating variants in *CDH1* (observed in 53% of the ILBCs) and *PIK3CA* (observed in 47% of the ILBCs), were the only somatic alterations reported in the context of ILBC (Pereira *et al.*, 2016). Mutational processes active in breast tumours were investigated via a somatic whole-genome sequencing study that involved 560 breast cancer samples, including 32/560 (6%) of classic ILBCs and 6/560 (1%) of pleomorphic ILBC cases. Although ILBC represented 7% of the sample size, the study was not focused on pointing out the histological differences in mutational processes (Nik-Zainal *et al.*, 2016).

In recent years, a number of studies have investigated the somatic mutation profile of ILBC tumours and reported the frequently mutated genes (Ciriello *et al.*, 2015; Desmedt *et al.*, 2016; Michaut *et al.*, 2016). By comparing the somatic genetic alterations in IDBC and ILBC, TCGA investigators have presented the genetic events specific to ILBC initiation and progression and the differentiating genetic events between ILBC and IDBC has also been highlighted in this study (Ciriello *et al.*, 2015). The most recurrent somatic variants and copy number alterations observed in ILBC are summarised in Table 1.1 and reviewed in detail in the following sections.

Table 1.1: Recurrent* somatic mutations and copy number alterations reported in ILBC.

Study	**(Ciriello et al., 2015)**	**(Desmedt *et al.*, 2016)**	**(Michaut *et al.*, 2016)**	**(Zhu *et al.*, 2018)**	**(Rosa-Rosa *et al.*, 2019)**	**(L Cao *et al.*, 2019)**
Method	WES, RNA-seq, Expression arrays, DNA methylation, SNP arrays, Protein expression (RPPA)	Targeted sequencing (360 cancer-related genes), Genome-wide CNA	Targeted sequencing (613 protein kinases and cancer related genes), CNA (SNP6 array), Gene expression (DNA microarrays), Protein expression (RPPA)	Targeted sequencing (1053 genes)	Targeted sequencing (34 genes), Targeted amplification (5 genes)	CNA (67 genes) using NanoString
Samples	ILBC (n=127) IDBC (n=490) mixed IDBC/ILBC (n=88)	ILBC (n=413)	ILBC (n=144)	Pleomorphic ILBC (n=17)	High grade ILBC with pleomorphic features (n=27)	ILBC (n=70)
Gene	**Somatic mutation frequency (% cases)**					
CDH1	63%	65%	43%	59%	89%	NR
PIK3CA	48%	43%	35%	53%	33%	NR
PTEN	8%	4%	1%	0%	NR	NR
TP53	8%	7%	4%	12%	19%	NR
TBX3	9%	13%	NR	24%	7%	NR
FOXA1	7%	9%	NR	6%	NR	NR
MAP3K1	6%	5%	5%	35%	19%	NR
AKT1	2%	4%	5%	6%	7%	NR
RUNX1	10%	3%	NR	12%	NR	NR
GATA3	5%	7%	5%	6%	7%	NR

ERBB2	4%	5%	4%	18%	26%	NR
ERBB3	n. s	4%	3%	24%	NR	NR
KMT2C	n. s	8%	NR	35%	19%	NR
ARID1A	n. s	6%	NR	6%	15%	NR
BRCA2	n. s	2%	4%	0%	4%	NR
Gene	**Frequent copy number alterations (% cases)**					
ESR1 amp[a]	NR	25%	NR	NR	NR	24%
ERBB2 amp[a]	7%	0%	4%	6%	4%	19%
CDH1 del[b]	89%	94%	NR	NR	NR	NR
MDM4 amp[a]	NR	NR	NR	NR	NR	17%
ARID1A del[b]	NR	23%	NR	NR	NR	NR
CCND1 amp[a]	17%	38%	15%	12%	11%	33%
MYC amp[a]	NR	31%	NR	NR	NR	17%
IGF1R amp[a]	NR	31%	NR	NR	NR	n. s
FGFR1 amp[a]	9%	25%	8%	NR	7%	n. s
TBX3	NR	19%	NR	NR	NR	NR
PTK2 del[b]	NR	18%	NR	NR	NR	NR

WES: Whole exome-sequencing. RNA-seq: RNA sequencing. RPPA: Reverse phase protein assay. CNA: Copy number alteration. ILBC: Invasive lobular breast cancer. IDBC: Invasive ductal breast cancer. [a] Amplification. [b] Deletion. n.s: not significant. NR: Not reported. * Recurrent somatic mutations and CNAs were determined using different methods in different studies. MutSigCV2 for somatic mutation and GISTIC for CNAs was used by (Ciriello et al., 2015); Somatic mutation obsereved in at least 2% of the cases and CNAs in at least 5% of the cases were defined as recurrent alterations by (Desmedt *et al.*, 2016); Binomial test based *P*-value for somatic mutation and ADMIRE tool for CNAs was used by (Michaut *et al.*, 2016); (Zhu *et al.*, 2018) and (Rosa-Rosa *et al.*, 2019) reported the total somatic mutations and CNAs; CN>= 5 (amplification), CN<5 (no amplification) were the cut-offs used by (L Cao *et al.*, 2019).

1.5.4.1 Loss of e-cadherin (*CDH1*) is the hallmark of ILBC

E-cadherin is a transmembrane glycoprotein, encoded by the *CDH1* gene that mediates cell-cell adhesion in epithelial tissues (Takeichi, 1991). A complete loss of e-cadherin protein expression has been observed in ~90% of ILBC tumours (Reed *et al.*, 2015).

CDH1 has been reported to be the most recurrently mutated gene in ILBC, with studies reporting 50-65% of ILBC cases harbouring somatic variants in this gene (Ciriello *et al.*, 2015; Desmedt *et al.*, 2016; Michaut *et al.*, 2016) (Table 1.1). These studies also reported a loss of heterozygosity (LOH) at *CDH1* in nearly all the cases. Comparing the somatic profiles of luminal A ILBC (n=106) and luminal A IDBC (n=201), TCGA investigators found a significant enrichment for *CDH1* variants in luminal A ILBC (63% ILBC *versus* 2% IDBC, MutSigCV2, q-value = $1.4x10^{-30}$) (Ciriello *et al.*, 2015). These somatic variants were found to be uniformly distributed along the *CDH1* coding region and were mostly (83%) truncating (Ciriello *et al.*, 2015). A heterozygous loss of *CDH1* was observed in 89% of the ILBC cases and it was associated with a downregulation of *CDH1* mRNA and protein expression levels (Ciriello *et al.*, 2015).

Michaut et al., (2016) reported somatic variants in *CDH1* in 43% of 144 ILBC cases and a significant reduction in mRNA (Wilcoxon, *P*-value = $2.4x10^{-5}$) and protein (Wilcoxon, *P*-value = $8.9x10^{-4}$) levels were also reported in *CDH1* mutant samples. However, reduction in mRNA and protein expression levels were also observed in some ILBC cases that had no somatic inactivating variants in *CDH1* (Michaut *et al.*, 2016). An insensitive variant detection that led to a false negative finding could be one of the explanations in this case. Other possibility could be the existence of other mechanisms than somatic genetic alterations for *CDH1* silencing including promoter

hypermethylation as proposed before (Droufakou *et al.*, 2001). However, DNA methylation changes were not reported in this study to confirm this hypothesis (Michaut *et al.*, 2016). While previous studies have shown promoter hypermethylation associated silencing of *CDH1* in ILBC tumours (Droufakou *et al.*, 2001; Sarrió *et al.*, 2003), recent methylation profiling by TCGA investigators did not identify *CDH1* promoter methylation in any of the ILBC cases in their study (Ciriello *et al.*, 2015).

Despite being the most commonly mutated gene in ILBC, the role of *CDH1* alterations alone as an ILBC driving event is not convincing. Derksen et al., (2006) using conditional *CDH1* gene inactivation in a mouse tumour model demonstrated that combined loss of e-cadherin and p53, but not e-cadherin alone resulted in tumour formation in the mice (Derksen *et al.*, 2006). In agreement with this, Pereira et al., (2016), investigating the somatic mutation profiles of 2,433 breast cancers including 194 ILBC cases, found a significant pattern of co-mutations in women with breast cancer between *CDH1* and other genes such as *PIK3CA* (OR = 2.1, 95% CI: 1.6-2.9), *TBX3* (OR = 3.2, 95% CI: 1.7-5.7), *RUNX1* (OR = 3.3, 95% CI: 1.5-6.6) and *ERBB2* (OR = 5.7, 95% CI: 2.7-12), suggesting that *CDH1* inactivation together with other genes may play a role in ILBC initiation and development (Pereira *et al.*, 2016).

Loss of e-cadherin may be associated with the characteristic discohesive growth pattern of ILBC considering the important role of e-cadherin in cell-cell adhesion. It has also been postulated that *CDH1* loss may have a role in the peculiar metastatic pattern of ILBC again due to its crucial role of cell adhesion (Desmedt *et al.*, 2017). Desmedt *et al.*, (2016) found an enrichment of *CDH1* variants in multifocal lobular cases compared with

unifocal, suggesting that the loss of e-cadherin might be linked to a tendency of discohesive cells to spread in the breast stroma (Desmedt *et al.*, 2016).

1.5.4.2 Mutations in key genes of PI3K-AKT signalling pathway

The key genes (*PIK3CA, PTEN* and *AKT1*) in the phosphatidylinositol 3-kinase (PI3K)/Akt pathway have been reported to be commonly mutated in ILBC (Ciriello *et al.*, 2015; Desmedt *et al.*, 2016; Michaut *et al.*, 2016). The PI3K/Akt pathway regulates many cellular functions related to cell growth, proliferation and survival (Bader *et al.*, 2005).

Somatic variants in *PIK3CA* represent the second most frequent somatic alterations in ILBC after *CDH1,* reported in 30-50% of all cases (Ciriello *et al.*, 2015; Desmedt *et al.*, 2016; Michaut *et al.*, 2016) (Table 1.1). Somatic variants targeting *PIK3CA* were found to be mainly missense variants and commonly associated with less proliferative tumours (defined by ki-67) (Desmedt *et al.*, 2016). Somatic alterations at *PTEN* (LOH and somatic variants) have been reported in up to 14% of all ILBC cases, whereas somatic variants in *AKT1* have been reported in up to 3% of all cases in different studies (Ciriello *et al.*, 2015; Desmedt *et al.*, 2016). An association between somatic variants at *AKT1* and increased risk of early relapse has been reported (Desmedt *et al.*, 2016). Notably, somatic alteration at *PIK3CA, PTEN* and *AKT1* were reported to be mutually exclusive to each other, suggesting that each of these genes may be supporting tumour growth and progression independently (Ciriello *et al.*, 2015; Desmedt *et al.*, 2016).

TCGA investigators reported a significantly higher mutation frequency at *PIK3CA* in ILBC cases in comparison with IDBC (48% of 127 ILBCs *versus* 33% of 490 IDBCs, MutSigCV2, q-value = 0.02). However, this difference was not significant when luminal A ILBC and luminal A IDBC were compared (Ciriello *et al.*, 2015). Inactivation

of *PTEN* (by homozygous deletion and somatic pathogenic variants) has been reported as a strong discriminating feature, after *CDH1* inactivation, between luminal A ILBC and luminal A IDBC, with a significantly higher frequency of *PTEN* inactivation observed in luminal A ILBC (14% of 106 luminal A ILBCs *versus* 3% of 201 luminal A IDBCs, MutSigCV2, q-value = $9x10^{4}$) (Ciriello *et al.*, 2015). A significant reduction (*P*-value = $4x10^{-4}$) in PTEN protein expression was also reported in luminal A ILBCs, as measured by reverse-phase protein assay (Ciriello *et al.*, 2015). Together, these somatic alterations in *PIK3CA*, *PTEN* and *AKT1* led to a significant upregulation of PI3K/Akt pathway-related proteins and phospho-proteins and an activation of PI3K/Akt pathway in ILBC tumours was observed in 45% of the cases (Ciriello *et al.*, 2015).

Findings from these studies suggest that PI3K/Akt signalling is commonly disrupted in ILBC suggesting its possibly crucial role in ILBC development. Treatment options targeting PI3K/Akt signalling pathway have been an active area of clinical research and could be a possible treatment option for ILBC tumours (Hosford & Miller, 2014; Klarenbeek *et al.*, 2020).

1.5.4.3 Mutations in transcriptional regulators

FOXA1 and *GATA3* are key regulators involved in the transcription regulation activity of the ER. ER is the master transcriptional regulator in breast cancer and is encoded by the *ESR1* gene (Y Zheng *et al.*, 2016). As estrogen interacts with ERs, a transcription factor complex is formed that binds to specific sites on the DNA leading to gene activation (Hua *et al.*, 2018). *FOXA1* and *GATA3,* mediate the transcription factor complex formation and thus are critical in the transcription of ER regulated genes (Hurtado *et al.*, 2011).

Somatic variants in *FOXA1* have been reported in up to 9% of ILBC cases (Ciriello *et al.*, 2015; Desmedt *et al.*, 2016) (Table 1.1). Luminal A ILBC was found to be enriched for somatic variants in *FOXA1* (7% of 106 luminal A ILBCs *versus* 2% of 201 luminal A IDBCs, MutSigCV2, q-value = 0.08) in TCGA study (Ciriello *et al.*, 2015). In ILBC, somatic variants in *FOXA1* were found to be specifically localised in the fork head DNA binding and C terminus transactivation domains, whereas in IDBC the variants were found in other structural elements also, without any preference (Ciriello *et al.*, 2015). In contrast, somatic variants in *GATA3*, another key ER modulator, were significantly lower in luminal A ILBC (5% of 106 luminal A ILBCs *versus* 20% of 201 luminal A IDBCs, MutSigCV2, q-value = 0.003) (Ciriello *et al.*, 2015). A significant reduction in *GATA3* mRNA (*P*-value = 0.007) and protein expression (*P*-value = $2x10^{-4}$) levels were reported in luminal A ILBC compared with luminal A IDBC and was proposed as another discriminatory feature between ILBC and IDBC.

Other transcription regulators reported to be mutated in ILBC include *TBX3*, a T-box gene family of transcription factors and chromatin regulatory factors, *KMT2C* and *ARID1A*. Somatic variants in *TBX3* has been reported in 9-13% of ILBC cases (Ciriello *et al.*, 2015; Desmedt *et al.*, 2016) (Table 1.1). A significant enrichment for *TBX3* variants was reported in ILBC in comparison with IDBC (9% of 127 ILBCs *versus* 2% of 490 IDBCs, MutSigCV2, q-value = 0.003) and this difference remained significant when comparing luminal A ILBC and luminal A IDBC (q-value = 0.05) (Ciriello *et al.*, 2015). Somatic variants in *KMT2C* and *ARID1A* was observed in up to 8% and 6% of the ILBC cases, respectively (Ciriello *et al.*, 2015; Desmedt *et al.*, 2016) (Table 1.1).

1.5.4.4 Other recurrent mutations in ILBC

Somatic variants in *ERBB2* and *ERBB3* have been reported in up to 5% and 3% of ILBC cases, respectively (Ciriello *et al.*, 2015; Desmedt *et al.*, 2016; Michaut *et al.*, 2016) (Table 1.1). Variants in *ERBB2* were more commonly observed in high-grade ILBC or the non-classic ILBC histological subtype (Desmedt *et al.*, 2016; Zhu *et al.*, 2018; Rosa-Rosa *et al.*, 2019). A significantly higher frequency of *ERBB2* variants was reported in pleomorphic ILBC compared with classic ILBC (21% of 21 pleomorphic ILBCs *versus* 2% of 49 classic ILBCs, *P*-value = 0.013) (Lien *et al.*, 2015). In another study, Rosa-Rosa et al., (2019) reported *ERBB2* variants in 26% of 27 high-grade ILBCs with pleomorphic features.

It has been shown that *CDH1* mutated ILBC cases that also harbour variants in *ERBB2*, showed a significantly worse survival compared with *CDH1* mutated ILBC cases without *ERBB2* variants (Log-rank test, DFS, *P*-value = 0.003, overall survival *P*-value = 0.0003) (Ping *et al.*, 2016). This may suggest that *CDH1* and *ERBB2* may work synergistically to promote ILBC tumourigenesis as similarly shown for tumours with variants at *CDH1* and *PTEN* (Derksen *et al.*, 2006). A significant pairwise interaction (OR= 5.7, 95% CI: 2.7-12) between variants in *ERBB2* and *CDH1* has also been observed (Pereira *et al.*, 2016). Furthermore, relapsed cases of ILBCs have been found to be enriched for *ERBB2* genomic alterations, with a study reporting 86% of 22 relapsed ILBC cases harbouring at least one actionable alteration in *ERBB2* (somatic variant, gene fusion, amplification) suggesting a possible role of *ERBB2* in ILBC progression (Ross *et al.*, 2013). This study further confirmed the enrichment of *ERBB2* somatic variants or gene fusion in *CDH1*-mutated ILBC cases (23% of 22 *CDH1*-mutated ILBCs) compared with ILBC tumours harbouring no variants in *CDH1* (2% of 286 ILBCs with no *CDH1* variants, *P*-value = 0.0006) (Ross *et al.*, 2013).

These finding points towards an active functional interaction between *CDH1* and *ERBB2* in ILBC tumourigenesis. The high frequency of *ERBB2* alterations in high-grade ILBCs and relapsed ILBC cases suggests its potential role in providing a growth advantage to the tumour. Therapy options targeting *ERBB2/ERBB3* could benefit a subset of ILBC and needs further exploration (Ben-Baruch *et al.*, 2015; Bidard *et al.*, 2015).

1.5.4.5 Recurrent copy number alterations in ILBC

Although ILBC frequently harbour somatic variants at specific sets of genes, this tumour type is known to have typically less extensive chromosomal changes compared with IDBC (Reed *et al.*, 2015).

CDH1 deletion has been reported to be a common event in ILBC tumours. Desmedt et al., (2016) reported arm-level and focal alterations at *CDH1* in 94% of 170 ILBC cases and a loss of e-cadherin expression was confirmed in 85% of the subset of cases for which IHC staining was available (156/170, 92%). A more frequent reduction in e-cadherin expression was reported in the tumours harbouring both *CDH1* somatic inactivating variants and deletions compared with the tumours that retained at least one intact copy of the gene (*P*-value < 0.001) (Desmedt *et al.*, 2016). A similar finding has been reported in TCGA study where a heterozygous loss of *CDH1* locus (16q) was reported in 89% of 127 ILBC cases (Ciriello *et al.*, 2015).

Apart from *CDH1*, copy number gains at *ESR1* have been reported in up to 25% of ILBC cases (Desmedt *et al.*, 2016; L Cao *et al.*, 2019). Cao et al., (2019) observed *ESR1* gains or amplifications in 24% of 70 ILBC samples that also showed a significantly high mRNA expression levels compared with the samples with normal copy number (*P*-value= 0.001) (L Cao *et al.*, 2019). A significant enrichment of *ESR1* copy number gains

was observed in women who had subsequent recurrence compared with women that did not show recurrence (39% of 18 recurrent ILBCs *versus* 19% of 52 non-recurrent ILBCs, *P*-value = 0.047) (L Cao *et al.*, 2019).

Other common genomic alterations reported in ILBC include focal gains at *CCND1* (11q13.3) observed in (38% of 170 ILBCs), *MYC* (8q24.21) (31% of 170 ILBCs), *IGF1R* (15q26.3) (31% of 170 ILBCs), *FGFR1* (8p11.23) (25% of 170 ILBCs), and *TBX3* (12q24.21) (19% of 170 ILBCs) (Desmedt *et al.*, 2016). Focal losses were reported in *ARID1A* (1p36.22) (23% of 170 ILBCs) and *PTK2* (8q24.23) (18% of 170 ILBCs) (Desmedt *et al.*, 2016). A frequent amplification in *MDM4* was observed in 17% of 70 ILBC cases (L Cao *et al.*, 2019). *MDM4* and *CCND1* amplifications (33% of 70 ILBCs) were found to be significantly correlated with their respective mRNA expression levels (*MDM4, P*-value = 0.02 and *CCND1, P*-value = 0.0004), whereas *ERBB2* and *MYC* amplification (observed in 19% and 17% of 70 ILBCs, respectively) did not show a significant correlation with gene expression levels in ILBCs (*ERBB2, P*-value = 0.1 and *MYC, P*-value = 0.34) (L Cao *et al.*, 2019).

1.6 Epigenetic modifications and tumourigenesis

The concept of epigenetics was first proposed by Conrad Waddington in 1939 (Waddington & H, 1942). Epigenetics is defined as a reversible and mitotically heritable modification to the DNA that leads to changes in gene expression without changing the actual DNA sequence.

DNA methylation is one of the most extensively studied epigenetic events in human cancers. It is a post-replication chemical modification of the DNA where, a methyl group

is added to the 5 prime cytosine located adjacent to a guanosine in a CpG dinucleotide (Bird, 1992). Most CpG dinucleotides are distributed throughout the human genome within repeat elements and transposons, whereas some are present in short dense clusters (200 bp-1000 bp) referred as CpG islands (Bird, 1992). CpG islands are mostly located at the transcription start site of human genes (Baylin, 2005). Approximately, half of all genes across the genome have CpG islands proximal to their promoter regions (near 5 prime UTR and 1st exon) (Baylin, 2005). In some instances, CpG islands are also found in the gene body or 3 prime UTR region and these CpG islands, located in unusual sites, have been found to be more prone to methylation (Nguyen *et al.*, 2001).

During early embryogenesis, almost all methylation patterns from the paternal and maternal genomes are erased and a new pattern is established and maintained through cell divisions by the family of DNA methyltransferases (Smith *et al.*, 2012). In a healthy cell, the genome-wide DNA methylation pattern has a bimodal distribution, which means that most of the CpG dinucleotides (~75%) within the repetitive elements and transposons are methylated, whereas those associated with the CpG islands within the gene promoter regions (<10%) are largely unmethylated (Bird, 2002). This pattern often becomes reversed during tumourigenesis and a global hypomethylation event together with hypermethylation of many tumour suppressor gene promoters is considered as a hallmark of cancer (Baylln *et al.*, 1997). Global hypomethylation leads to genomic instability that promotes tumour progression (Kulis & Esteller, 2010). It has been shown that the levels of global hypomethylation progress throughout different stages of tumourigenesis (Fraga *et al.*, 2004). On average, a tumour cell looses 20-30% of methylation compared with the adjacent normal cells and the level of global methylation reduction also varies between different tumour types (Ehrlich, 2000; Wilson *et al.*, 2007). Breast cancer shows up to a

50% reduction in the total methylated CpG content compared to the adjacent normal epithelium (Wilson *et al.*, 2007). In the past decade, numerous studies have investigated promoter hypermethylation landscape and suggested that unique CpG island hypermethylation profiles exist for specific tumour types (Costello *et al.*, 2000; Esteller *et al.*, 2001; James G. Herman & Baylin, 2003).

1.7 DNA methylation in breast cancer

Several factors contribute to the pathogenesis of breast cancer including DNA methylation (Widschwendter & Jones, 2002). DNA Methylation has been shown to be an early event in breast carcinogenesis, resulting in the activation of many oncogenes and silencing of tumour-suppressor genes, thus providing a growth advantage to the tumour cells (Evron *et al.*, 2001; Widschwendter & Jones, 2002). There are several genes reported to be hypermethylated in breast cancer including genes involved in important cellular pathways such as cell cycle regulation (*CDKN2A*, *CCND2)* (Fackler *et al.*, 2004; Radpour *et al.*, 2009), DNA repair (*BRCA1, GSTP1*) (Radpour *et al.*, 2009), metastasis (*RASSF1A, RARβ2, TWIST* and *HIN1*) (Fackler *et al.*, 2004), cell adhesion (*CDH1*) (Caldeira *et al.*, 2006) and hormone-mediated cell signalling (*ESR1*, *PGR*) (Lapidus *et al.*, 1996). Genome-wide hypomethylation is also frequently observed in breast cancers, and many genes have been reported to be hypomethylated including *IL10* (Son *et al.*, 2010), *NAT1* (SJ Kim *et al.*, 2008), *MDR1* (G Sharma *et al.*, 2010), *FEN1* (Singh *et al.*, 2008), *CDH3* (Paredes *et al.*, 2005), *JAGGED1* and *NOTCH1* (Y Cao *et al.*, 2015).

DNA methylation has been associated with clinicopathological features of the tumour such as tumour grade and tumour stage. Yan et al., (2000) showed that CpG island hypermethylation is related with histological grade in breast tumours with high grade

tumours showing an increased number of methylated CpG islands (PS Yan *et al.*, 2000). Furthermore, different breast cancer subtypes have been shown to be associated with distinct DNA methylation pattern (Bediaga *et al.*, 2010; Holm *et al.*, 2010; Kamalakaran *et al.*, 2011; Stefansson *et al.*, 2015). Holm et al., (2010) using methylation levels of 807 cancer-related genes in 189 breast cancer samples and four normal breast samples showed that the intrinsic subtypes display distinct methylation patterns. At the differentially methylated CpGs that corresponded to 163 genes, luminal B tumours were found to be the most and basal-like tumours were found to be the least frequently methylated (*P*-value = $2x10^{-7}$). Lee et al., (2010) assessing the methylation status in a panel of 10 gene in 57 luminal, 24 HER2-enriched and 33 basal-like breast cancer showed that the median methylation levels of *HIN1, RASSF1A* and *TWIST*, and the average methylation ratio were significantly lower in basal-like subtype compared to luminal or HER2 subtypes. In contrast, *BRCA1* methylation level was significantly higher in basal-like subtype than in luminal subtype (JS Lee *et al.*, 2010; Stefansson *et al.*, 2015). This suggests that DNA methylation may have a role in the breast cancer heterogeneity and the development of distinct breast cancer subtypes.

1.7.1 DNA methylation as a biomarker for disease prognosis and treatment response

DNA methylation signatures have emerged as an important tool for prognosis prediction as well as predicting treatment response in cancer (Mikeska *et al.*, 2012; Hao *et al.*, 2017; Leygo *et al.*, 2017).

The global DNA methylation pattern or methylation at a specific gene in tumours can be indicative of patient prognosis. Methylation at many gene promoters has been

suggested to have independent prognostic value in breast cancer including *HOXA11* (Xia *et al.*, 2017), *ESR1* and *PITX2* (Sheng *et al.*, 2017), *HOXD13* (Zhong *et al.*, 2015) and *CDH22* (Martín-Sánchez *et al.*, 2017). *BRCA1* promoter methylation has been shown to be significantly correlated with poor overall survival in women with breast cancer (pooled HR=1.38, 95% CI: 1.04-1.84) (Wu *et al.*, 2013). Another study reported *RASSF1* promoter methylation to be significantly associated with poor prognosis (disease free survival, pooled HR = 2.54 (95% CI: 1.77-3.66) (Jiang *et al.*, 2012). *APC* and p16 promoter methylation have also been shown to predict disease outcome (X Xu *et al.*, 2010). Debouki-Joudi et al., (2017) evaluated *APC* promoter methylation in 91 sporadic and 44 familial breast cancer cases and found a similar frequency of *APC* promoter methylation across sporadic and familial breast cancer cases, (52% sporadic, and 54% familial cases). They found that in both sporadic and familial breast cancer cases, *APC* promoter hypermethylation was associated with aggressive tumour behaviour and poor survival (Debouki-Joudi *et al.*, 2017).

There has been a growing interest in identifying and utilising DNA methylation signatures to refine breast cancer molecular classification and improve the prognostic abilities in the clinical setting. Holm et al., (2010) showed that gene expression-based breast cancer subtypes had specific methylation profiles with luminal B and basal-like subtypes being the most and the least frequently methylated, respectively (Holm *et al.*, 2010). Another study, using DNA methylation levels at 3,869 CpG sites in 669 breast cancer samples, identified nine subgroups with significant difference in patient prognosis (Log-rank test, P-value < 0.004) (S Zhang *et al.*, 2018). They identified further subgroups within the basal-like subtype with distinct methylation profiles with a significant difference in survival (Log-rank test, P-value < 0.04). They suggested that the

methylation-based subgroups were more elaborate compared with the gene expression-based subtypes (S Zhang *et al.*, 2018). Fleischer et. al., (2017), based on a hierarchical clustering of luminal A breast tumours using a DNA methylation signature (SAM40, that included 41 differentially methylated genes), further segregated luminal A tumours into two subgroups and found that the subgroup with low relative methylation showed a significantly better prognosis compared with the subgroup that had high relative methylation (Log-rank test, *P*-value = 0.001) (Fleischer *et al.*, 2017). Another study, using data from TCGA, identified methylation sites including *SOSTDC1*, *ESCO2*, *CDCA2*, *PTN*, *RGMA*, *KLK4* and *CENPA* that showed prognostic value in luminal breast cancer (Xiao *et al.*, 2018). Specific to TNBCs, a study based on whole-genome methylation sequencing, stratified TNBCs into three methylation clusters with the hypomethylated cluster showing better prognosis compared with the other two highly methylated clusters (HR = 8.64, *P*-value = 0.005) (Stirzaker *et al.*, 2015).

Global changes in DNA methylation patterns and their association with breast cancer therapies holds a great potential in predicting treatment response and identifying more relevant treatment subgroups. A growing number of large-scale methylation studies are investigating this in breast cancer. Martens et al., (2005) investigated the promoter methylation status of 117 genes in 200 hormone receptor positive tumours in women who received the antiestrogen tamoxifen as first line of treatment for recurrent breast cancer and identified 10 genes for which promoter methylation was significantly correlated with clinical benefit to antiestrogen therapy with the strongest being for *PSAT1* (*P*-value < 0.0001) (Martens *et al.*, 2005). In tamoxifen-treated, hormone receptor positive, lymph node negative breast cancers, *PITX2* promoter methylation was associated with increased risk of recurrence (HR= 2.75, 95% CI: 1.40-5.41) (Harbeck *et al.*, 2008). *PITX2* promoter

methylation was associated with poor patient outcome in breast cancer patients treated with anthracycline-based adjuvant chemotherapy (HR=1.46, 95%CI: 1.05-2.01) (Nimmrich *et al.*, 2008). *BRCA1* promoter methylation has been shown to be an independent favourable predictor of disease free survival and disease specific survival in women with TNBC who received adjuvant chemotherapy (HR, disease free survival= 0.45, 95% CI: 0.24-0.84, *P*-value= 0.02; HR, disease specific survival = 0.43, 95% CI: 0.19-0.95, *P*-value= 0.04) (Y Xu *et al.*, 2013).

Taken together, these studies clearly demonstrate that DNA methylation signatures could be a valuable tool in clinical settings for refining breast cancer prognostication and management.

1.8 DNA methylation alterations in ILBC

Tumour DNA methylation has been shown to have great potential for refining breast cancer classification and also have predictive value for disease prognosis and therapy response (as reviewed in section 1.7.1). However, DNA methylation alterations associated with ILBC initiation and progression has not been studied adequately. Most studies focusing on ILBC specific methylation alterations have primarily investigated a set of specific candidate genes as summarised in Table 1.2 and data on genome-wide methylation changes and their role in ILBC tumourigenesis is limited.

Table 1.2: Previous studies investigating the DNA methylation pattern in ILBC.

Study	Gene promoter	Sample	Assay	Finding
(Droufakou *et al.*, 2001)	*CDH1*	ILBC (n=22)	Methylation specific PCR	- *CDH1* promoter methylation in 77% of ILBCs, of which 65% were negative for e-cadherin negative (measured by IHC).
(Lehmann *et al.*, 2002)	*DAPK1*	ILBC (n=19) IDBC (n=85)	Methylation specific PCR	- *DAPK1* promoter methylation in 53% of ILBCs and 9% of IDBCs.
(Sarrió *et al.*, 2003)	*CDH1*, *APC*, *CTNNB1*	ILBC (n=46)	Methylation specific PCR	- *CDH1* promoter methylation in 41% of ILBCs. - *APC* promoter methylation in 56% of ILBCs.
(Fackler *et al.*, 2003)	*RASSF1, HIN-1, RAR- β, Cyclin- D2, TWIST*	ILBC (n=19) IDBC (n=27) LCIS (n=13) DCIS (n=44)	Methylation specific PCR	- Similar methylation profiles of ILBCs and IDBCs with respect for *RASSF1*, *HIN1*, *RAR- β* and *Cyclin- D2*. - *TWIST1* promoter methylation in 16% of ILBCs and 56% of IDBCs.
(Bae *et al.*, 2004)	12 genes- RAR- *β, Cyclin- D2, TWIST, ER, CDH1, BRCA1, THRβ, GSTP1, HIN-1, RASSF1A, BAX, RB*	ILBC (n=19) IDBC (n=60) mucinous breast cancer (n=30)	Methylation specific PCR	- ILBCs and mucinous breast cancer samples showed relatively higher frequencies of methylation compared with IDBC (49% in ILBCs and mucinous *versus* 40% in IDBCs). - *BRCA1* promoter methylation in 92% of mucinous breast cancer, 39% of ILBCs and 28% of IDBCs.
(Lo *et al.*, 2006)	*SFRP1*	ILBC (n=9) IDBC (n=28) LCIS (n=9) DCIS (n=19) Normal breast (n=14)	Methylation specific PCR	- *SFRP1* promoter methylation in 33% of ILBCs and 68% of IDBCs.
(Caldeira *et al.*, 2006)	*CDH1*	ILBC (n=5) IDBC (n=71) other (n=3)	Methylation specific PCR	- *CDH1* promoter methylation in 80% of ILBCs and 73% of IDBCs.
(Seniski *et al.*, 2009)	*ADAM33*	ILBC (n=21) IDBC (n=51)	Methylation-specific PCR	- *ADAM33* promoter methylation in 76% ILBCs and 26% of IDBCs.

(Zou *et al.*, 2009a)	*CDH1*	ILBC (n=14) LCIS (n=13)	Methylation-specific PCR	- *CDH1* promoter methylation in 93% of ILBCs and all LCIS samples.
(Tserga *et al.*, 2012)	*DCR1, DAPK1, RASSF1A, DCR2, APC, MGMT, GSTP1 and PTEN*	ILBC (n=9) IDBC (n=34) Mixed lobular-ductal carcinoma (n=2)	Methylation-specific PCR, Methylation-sensitive high-resolution melting analysis	- *MGMT* promoter methylation in 11% of ILBC and 24% of IDBCs - *GSTP1* promoter methylation in 11% of ILBCs and 21% of IDBCs - *PTEN* promoter methylation in 11% of ILBCs and 6% of IDBCs - *APC* promoter methylation in 75% of ILBCs and 55% of IDBCs - *RASSF1* promoter methylation in 38% of ILBCs and 39% of IDBCs - *DAPK1* promoter methylation in 50% of ILBCs and 39% of IDBCs - *DCR1* promoter methylation in 63% of ILBCs and 35% of IDBCs
(Medina-Jaime *et al.*, 2014)	*ESR1*, *PGR*	ILBC (n=20) IDBC (n=20)	Methylation-specific PCR	- Promoter methylation at *ESR1* and *PGR* were found to be uncommon in ILBC and IDBC.
(Moelans *et al.*, 2015)	24 putative tumour suppressor genes	classic-ILBC (n=20) pleomorphic ILBC (n=16) IDBC (n=20).	MS-MLPA	- Low *TP73* and *MLH1* promoter methylation and relatively high *RASSF1A* promoter methylation levels in pleomorphic ILBC compared with classic ILBC. - Low *MLH1* and *BRCA1* methylation levels in pleomorphic ILBC compared with IDBC - Pleomorphic ILBC and IDBC showed similar methylation patterns, while the methylation pattern of classic ILC was different.
(Ciriello *et al.*, 2015)	Genome-wide DNA methylation	ILBC (n=201)	HM450K	- No methylation at *CDH1* gene promoter. - Promoter methylation at *FOXA1* that correlated with reduction in gene expression.
(Roessler *et al.*, 2015)	Genome-wide DNA methylation	ILBC (n=10) sporadic IDBC (n=10) IDBC with germline BRCA1 mutation (n=8) Normal Breast tissue (n=4)	HM450K	- ILBCs showed the highest methylation among the breast cancer types and had the lowest genetic instability (the number of copy number alteration, as measured by array-based comparative genomic hybridisation) - IDBCs with germline *BRCA1* mutation showed the lowest levels of DNA methylation but the highest levels of genetic instability. - Strong evidence to support the existence of CpG Island Methylation Phenotype associated specifically with the ILBC type.

ILBC: Invasive lobular breast cancer. IDBC: Invasive ductal breast cancer. LCIS: Lobular carcinoma in-situ. DCIS: Ductal carcinoma in-situ. HM450K: Illumina HumanMethylation 450K BeadChip array. MS-MLPA: Methylation-specific multiplex ligation-dependent probe amplification.

Roessler *et al*. were the first study to perform genome-wide DNA methylation profiling of ILBC. They profiled ten ILBCs, ten sporadic IDBCs, ten IDBCs with a germline *BRCA1* mutation and four normal breast tissues and investigated the existence of CpG island methylator phenotype in breast tumours. Based on the methylation level of seven genes (*DNM3*, *mir129-2*, *PGLYRP2*, *PRKCB*, *RGS7*, *SHF* and *TACC1*), they found that ILBC cases formed a distinct hypermethylation cluster with a subset of sporadic IDBCs while the *BRCA1*-mutated IDBC samples were found to cluster with the normal breast tissue. They also demonstrated that aberrant DNA hypermethylation was negatively correlated with the copy number alterations (as measured using array-based comparative genomic hybridisation), with *BRCA1*-mutated IDBC tumours harbouring a higher number of genomic alterations (mean = 125) compared with the ILBC tumours (mean = 11). This study suggested a strong association of CpG island methylator phenotype with the lobular phenotype. However, no further ILBC-associated DNA methylation alterations were reported (Roessler *et al.*, 2015).

Considering the high frequency of *CDH1* somatic mutations in ILBC, *CDH1* promoter methylation has been most extensively investigated in ILBC (Droufakou *et al.*, 2001; Sarrió *et al.*, 2003; Bae *et al.*, 2004; Caldeira *et al.*, 2006; Zou *et al.*, 2009b). Zou *et al*., (2009) demonstrated that *CDH1* promoter hypermethylation is an early event in ILBC tumourigenesis as it was observed in all 13 LCIS cases in the study, whereas 93% of 14 ILBC cases showed *CDH1* promoter methylation (Zou *et al.*, 2009a). Droufakou *et al*., (2001) reported *CDH1* promoter methylation in 77% of 22 ILBC cases, of which 65% showed downregulation of e-cadherin protein (as measured by IHC). Caldeira et al., (2006) reported *CDH1* promoter methylation in 73% of 77 IDBC samples and 80% of 5

ILBC samples. This suggests that *CDH1* promoter methylation may not be an exclusive event in ILBC tumourigenesis.

Several studies have reported genes that showed differential methylation between ILBC and other breast cancer subtypes (Table 1.2). *TWIST*, a transcription factor involved in early development, was found to be less frequently hypermethylated in ILBCs compared with IDBCs (16% of 19 ILBCs versus 56% of 27 IDBCs, *P*-value =0.01) (Fackler *et al.*, 2003). A significant difference in *DAPK1* promoter methylation between ILBC and IDBC has been reported in another study where they noted promoter hypermethylation in 53% of 19 ILBC samples compared with 9% of 85 IDBC samples, *P*-value < 0.001) (Lehmann *et al.*, 2002). They also found that *DAPK1* promoter hypermethylation significantly correlated with loss of mRNA expression (as measured by quantitative RT-qPCR), ER positive tumours and absence of *TP53* overexpression (*P*-value < 0.001). They suggested a possible role of *DAPK1* in tumour progression as no promoter hypermethylation was observed in the LCIS samples (Lehmann *et al.*, 2002). Furthermore, promoter hypermethylation of *ADAM33* has been shown to be more frequently methylated in ILBCs compared with IDBCs (76% of 21 ILBCs *versus* 26% of 51 IDBCs, *P*-value= 0.0002) and a reduction in gene expression was also observed in hypermethylated samples (using RT-PCR) (Seniski *et al.*, 2009). In an attempt to investigate the DNA methylation profiles of different breast cancer histological types, Bae et. al., (2004) compared the DNA methylation profiles of ILBC, IDBC and mucinous breast cancer samples across a set of 12 genes and found that ILBC and mucinous breast cancer samples showed relatively higher frequencies of hypermethylation compared with IDBC samples (49% of 19 ILBCs, 49% of 30 mucinous breast cancer *versus* 40% of 60 IDBCs). They also found that *BRCA1* showed significantly different methylation

frequencies between the breast cancer subtypes with 92% of mucinous breast cancer showing *BRCA1* promoter hypermethylation compared with 39% of ILBC and 28% of IDBC samples (P <0.001) (Bae *et al.*, 2004).

Recently, TCGA profiled 201 ILBC tumours at different molecular levels that also included genome-wide DNA methylation. However, the analysis of the study mainly focused on the somatic mutation profile of ILBC. Recurrently altered genes in ILBC were discussed in detail while little emphasis was given on the methylation data. They only reported the promoter methylation status at two genes, *CDH1* and *FOXA1*. Whilst no *CDH1* promoter hypermethylation was reported in ILBC samples, *FOXA1* showed promoter methylation at the binding site and it was found to be negatively correlated with the gene expression (Ciriello *et al.*, 2015).

These findings suggest that ILBCs display a different DNA methylation profile compared to other breast cancer types across several genes that could impact gene expression. One common limitation of most of the above-reviewed studies is the small sample size that makes them underpowered. Additionally, many of these studies have been based on candidate gene approaches and have investigated methylation status at a specific gene promoter or a panel of gene promoters. This limits the analysis to confined regions of the genome. A genome-wide analysis of DNA methylation alterations may potentially reveal new genes and functional gene networks associated with ILBC tumourigenesis.

1.9 Statement of problem, hypothesis and aims

This review has discussed ILBC as a distinct and heterogeneous breast cancer subtype, focusing on its clinical behaviour and the key genetic and epigenetic (DNA methylation) alterations in ILBC tumourigenesis and progression. Recent studies have shown that ILBC has distinct genomic features compared with IDBC. Somatic genetic alterations (somatic variants and LOH) targeting *CDH1* have been recognised as the most frequent and prevalent alterations in ILBC tumours with ~90% of ILBC showing a complete loss of e-cadherin. Somatic variants in other genes such as *PIK3CA* (a key gene in the PI3K/Akt signalling pathway), *FOXA1* and *TBX3* have also been more commonly observed in ILBC tumours compared with IDBC. Inactivating genomic alterations including somatic variants and homozygous losses of the *PTEN* locus are another discriminating feature between ILBC and IDBC being more frequent in ILBC. ILBC also shows within-subtype heterogeneity with somatic variants in genes such as *TP53*, *HER2* and *HER3* being more commonly observed in the solid and pleomorphic ILBC histological subtypes.

Current breast cancer studies have demonstrated the existence of further subgroups within the gene expression-based subtypes with significant difference in prognosis. Tumour DNA methylation has been shown to have great potential for further refining breast cancer classification. However, data on DNA methylation alterations specific to ILBC initiation and progression is limited.

Despite having a distinct clinical behaviour, molecular profile, patient outcome and response to therapy, ILBC does not yet have any specific treatment regimen. Investigation of genome-wide DNA methylation may further improve our understanding of ILBC. A

combined genomic (somatic whole-exome sequencing) and epigenomic (DNA methylation) approach may help refine the ILBC heterogeneity and identify subgroup-specific prognostic markers and therapeutic targets that could improve precision medicine for ILBC.

The hypotheses of this study are:

1. ILBC tumours have a distinct genome-wide DNA methylation profile as compared to non-ILBC tumours.
2. Genome-wide variation in DNA methylation patterns within ILBC reflect different tumour biologies and can be used as a prognostic biomarker.
3. Homogeneous subgroups of ILBC may be identified using genetic and epigenetic data.

These hypotheses have been addressed by the three main aims of this study.

Aim I (Chapter 3): To investigate the genome-wide DNA methylation profiles of ILBC.

Sub-aim i: To identify the differentially methylated regions between ILBC and non-ILBC.

Sub-aim ii: To identify the variably methylated regions across the ILBC methylome.

Sub-aim iii: To assess the association between tumour methylation at the most variably methylated regions and overall survival for ILBC women.

Aim II (Chapter 4): To use genome-wide DNA methylation data to subgroup ILBC.

Aim III (Chapter 5): To further characterise the identified subgroups using somatic whole-exome sequencing data.

Chapter 2 Materials and Methods

2.1 Study participants

This study included 502 invasive breast cancer cases classified as invasive lobular breast cancers (ILBCs, n=161) and non-lobular invasive breast cancers (non-ILBCs, n=341), based on the International Classification of Diseases for Oncology (ICD-O) code (World Health Organization *et al.*, 2000) registered for each case on the Victorian Cancer Registry, Cancer Council Victoria (Victorian Cancer Registry, 2020) and confirmed by expert Pathological review. The non-ILBC cases in this study were predominantly (91%) invasive ductal breast cancers (IDBCs), ICD-O code-8500. The clinical and pathological features of the study participants are summarised in Table 2.1. Matching adjacent normal breast tissue samples were available for 13 breast cancer cases as previously described in (Wong *et al.*, 2016).

Table 2.1: Clinical and pathological features of the study participants.

Sample characteristics	ILBC (n=151)	non-ILBC (n=341)	*P*-value*
Median age at cancer diagnosis, years [interquartile range]	65 [25%; 57]	64 [25%; 58]	0.99
Age group, n (%)			
<50	12 (8)	16 (5)	0.29
50-60	41 (27)	91 (27)	
60+	94 (62)	233 (68)	
Missing[†]	4 (3)	1 (0.3)	
Year of diagnosis, n (%)			
1992-1996	22 (15)	66 (19)	$9.1x10^{-6}$
1997-2001	48 (32)	132 (39)	
2001-2006	40 (27)	110 (32)	
2006 and later	41 (27)	32 (9)	
Missing	0 (0)	1 (0.3)	
Ethnicity, n (%)			
Australian/NZ	112 (74)	269 (79)	0.30
Greek	4 (3)	17 (5)	
Italian	10 (7)	32 (9)	
UK/Malta	4 (3)	22 (6)	
Missing	21 (14)	1 (0.3)	
Study, n (%)			
MCCS[a]	130 (86)	341 (100)	$1.7x10^{-11}$
kConFab[b]	6 (4)	0 (0)	
ABCFR[c]	15 (10)	0 (0)	
Tumour ICD-O code[d], n (%)			
8500	1 (1)	312 (91)	$1.6x10^{-87}$
8520	95 (63)	6 (2)	
8522	34 (23)	2 (0.6)	
8211	0 (0)	11 (3)	
other	0 (0)	10 (3)	
Missing	21 (14)	0 (0)	
Tumour Grade, n (%)			
Grade I	14 (9)	77 (23)	$4.2x10^{-10}$
Grade II	92 (61)	127 (37)	
Grade III	19 (13)	121 (36)	
Missing	26 (17)	16 (5)	
Tumour Stage, n (%)			
1A/1B	65 (43)	Missing	-
2A/2B	48 (32)	Missing	
3A/3C/4	17 (11)	Missing	
Missing	21 (14)	Missing	

Median tumour purity, % [interquartile range]	58 [25%, 53]	61 [25%, 53]	0.13
Tumour ER expression, n (%)			
Positive	135 (89)	247 (72)	$1.3x10^{-6}$
Negative	10 (7)	88 (26)	
Missing	6 (4)	6 (2)	
Tumour PR expression, n (%)			
Positive	106 (70)	170 (50)	$3.4x10^{-6}$
Negative	38 (25)	165 (48)	
Missing	7 (5)	6 (2)	
Tumour HER2 expression, n (%)			
Positive	15 (10)	103 (30)	$8.6x10^{-7}$
Negative	101 (67)	231 (68)	
Equivocal	5 (3)	0 (0)	
Missing	30 (20)	7 (2)	
Molecular subtype[e], n (%)			
Luminal A	98 (65)	185 (54)	$5.8x10^{-8}$
Luminal B	13 (9)	76 (22)	
HER2 type	1 (0.6)	28 (8)	
Triple negative	3 (2)	50 (15)	
Missing	36 (24)	2 (0.5)	

ILBC: Invasive lobular breast cancer. non-ILBC: non-lobular invasive breast cancer. * *P*-values are for chi-square test and T.test for categorical and continuous variables, respectively. † Data not available [a] The Melbourne Collaborative Cohort Study. [b] The Kathleen Cuningham Foundation Consortium for Research into Familial Breast Cancer. [c] Australian Breast Cancer Family Registry. [d] International Classification of Diseases for Oncology cancer code (8500-Invasive ductal breast cancer; 8520-Invasive lobular breast cancer; 8522-Infiltrating ductal and lobular carcinoma; 8211-Tubular carcinoma). ER: Estrogen receptor, PR: Progesterone receptor, HER2: Human epidermal growth factor receptor 2. The ER, PR and HER2 expression status were measured using immunohistochemistry as described (Blows *et al.*, 2010).[e] Molecular subtypes were defined using definition reported by the St Gallen International Expert Consensus as: Luminal A: ER and/or PR positive and HER2 negative; Luminal B: ER and/or PR positive and HER2 positive; HER2 type: ER and PR negative and HER2 positive and Triple negative: ER negative, PR negative and HER 2 negative (Goldhirsch *et al.*, 2011).

The breast cancer samples were sourced from three well-characterised Australian studies: Melbourne Collaborative Cohort Study (MCCS) (R Milne *et al.*, 2017), Australian Breast Cancer Family Registry (ABCFR) (John *et al.*, 2004) and The Kathleen Cuningham Foundation Consortium for Research into Familial Breast Cancer (kConFab) (Mann *et al.*, 2006). Study designs and the data collection methods of the three studies are presented in detail in the sections below.

2.1.1 The Melbourne Collaborative Cohort Study

The Melbourne Collaborative Cohort Study (MCCS) is one of the largest prospective cohort studies conducted in Australia (R Milne *et al.*, 2017). It was set up in 1990 to prospectively investigate the role of diet and other lifestyle factors in the predisposition of cancer mainly in prostate cancer, breast cancer and bowel cancer. Between 1990 and 1994, 41,513 people aged between 40-69 years were recruited to this study. All participants were of white European origin; 69% born in Australia or New Zealand, 13% born in Italy, 11% in Greece and 6% in the UK.

Baseline information in the MCCS was collected by interviewer-administered questionnaires on lifestyle, personal medical history and medications taken. A self-administered questionnaire on diet was also filled by the participants. Blood samples were collected from 41,133 participants. Plasma, peripheral blood mononuclear cells and buffy coats were separated and stored in liquid nitrogen. From the second year of recruitment, dried blood spots were stored on Guthrie Cards (GCs). Formalin-fixed paraffin-embedded (FFPE) tissue was collected from 3,070 tumours diagnosed in cohort participants, the majority being breast, colorectal and prostate cancers (R Milne *et al.*, 2017).

The first follow-up occurred between 1995 and 1998, around four years after recruitment where information on lifestyle, medical history and diet was collected. The second follow-up occurred between 2003 and 2007 where further blood samples were collected, and repeated measures of key lifestyle exposures were recorded. Since 2005, the participants are contacted every year for an update. Incidences of cancer cases and mortality were updated regularly by matching the study to the Victorian Cancer Registry, Cancer Council Victoria and death indices.

2.1.2 The Kathleen Cuningham Foundation Consortium for Research into Familial Breast Cancer

The Kathleen Cuningham Foundation Consortium for Research into Familial Breast Cancer (abbreviated to "kConFab") is a familial breast cancer consortium, which was established in 1997 (Mann *et al.*, 2006). Families with a strong history of breast and breast-ovarian cancers were recruited to the kConFab through Family Cancer Centres in Australia and New Zealand.

As of 2018, 1,848 families were recruited to kConFab including 1,074 families with a strong history of breast cancer. All cancer cases reported in the family were verified with clinical pathology reports. Biological specimens (blood, normal and tumour tissues), family history, epidemiological, clinical and psychosocial data were collected from affected and unaffected female and male participants over 18 years of age. The blood samples collected were returned to the central core laboratory and processed following the standard protocol (kConFab biospecimen protocol). The normal and tumour tissue specimen were collected after surgery and dissected into 3 mm sections by a clinical pathologist. After initial assessment of the specimen, five 10 μm thick tissue sections

were stored in a cryovial for each sample in liquid nitrogen at the kConFab tissue bank. Details about the biological specimens and clinical data were stored in a de-identified central relational database which has been made available to peer-reviewed, ethically approved, and funded research projects both nationally and internationally.

2.1.3 The Breast Cancer Family Registry

The Breast Cancer Family Registry was established in 1995 by the National Cancer Institute (NCI, USA) (John *et al.*, 2004). The participating institutes were from the USA, Canada and Australia and the families were recruited either directly by the cancer registries (population-based families) or from the clinics (clinic-based families).

Australian Breast Cancer Family Registry (ABCFR) is a component of the BCFR carried out in Melbourne and Sydney, Australia (Hopper *et al.*, 1994; McCredie *et al.*, 1998). It is a population-based case-control family study of breast cancer with an emphasis on early-onset disease. In this study, all adult women living in metropolitan Melbourne and Sydney who were diagnosed with a histologically confirmed primary breast cancer were invited to participate. Cases were identified by the use of the Victorian and New South Wales cancer registries, to which notification of cancer diagnosis is a legislative requirement. From 1992 to 1999 in Melbourne and from 1993 to 1998 in Sydney, women younger than 40 years at diagnosis were recruited to this study; after 1996, random samples of women aged 40-49 years and 50-59 years at diagnosis were also selected. Eligible women were recruited irrespective of family history of breast cancer.

Unaffected control subjects were randomly selected from general population living in Melbourne and Sydney using the electoral poll. A risk factor and family cancer

history questionnaire involving all known first- and second-degree relatives were completed by each participating case patients and control subjects. Blood and tumour specimen were collected from each affected woman. A risk factor and family cancer history questionnaire involving all known first- and second-degree relatives were completed by all participating case patients and control subjects.

2.2 Study governance and data acquisition

This project was formally approved via an application to the governance groups establish by each of the research resources included in these analyses. Clinical, pathological and epidemiological data was requested using a Data Request Form provided by each research resource. The following information was requested and made available: i) tumour details ii) breast cancer treatment and outcome data iii) immunohistochemical data related to the tumour iv) family history and v) lifestyle data. The data were obtained in a spreadsheet and a data dictionary was provided explaining the data fields. Data that were not available were marked as “Missing”.

2.3 DNA extraction

2.3.1 Formalin-fixed paraffin-embedded tumour tissue

Tumour enriched DNA was prepared from formalin-fixed paraffin-embedded (FFPE) tumour tissue sections using the QIAamp DNA FFPE Tissue kit (Qiagen, Germany) following the manufacturer’s instructions. Areas enriched with tumour cells were marked up by a trained pathologist on the hematoxylin and eosin-stained slides by matching the section on the FFPE slide (10 μm thick). To remove the paraffin wax from the tumour tissues, the FFPE slides were repeatedly washed three times each with xylene

and 100% ethanol. The slides were washed briefly in 2 changes of distilled water and stained using 0.1% methyl green solution.

The tumour-enriched area was macrodissected using a 21G needle (Terumo, Japan) and transferred to a 1.5 mL microfuge tube (Eppendorf, Germany) to which 180 µl of ATL lysis buffer (Qiagen, Germany) and 20 µl of proteinase K (Sigma-Aldrich, Germany) were added and vortexed thoroughly. The mixture was incubated at 56°C for 48 hours with intermittent vortex mixing. After incubation, the mixture was vortexed, briefly centrifuged and incubated at 90°C for 55 minutes. To the mixture, 200 µl of AL buffer and 200 µl of 100% ethanol were added and vortexed thoroughly to precipitate the DNA. The entire lysate was transferred to a QIAamp MinElute column (Qiagen, Germany) placed in a 2 ml collection tube (Qiagen, Germany) and centrifuged at 6000 relative centrifugal force (rcf) for 1 minute. The flow through was discarded and DNA was washed using 500 µl of buffer AW1. Another wash was performed using 500 µl of AW2 buffer. The column was centrifuged at full speed for 3 minutes to completely dry the membrane. A triple elution was performed to elute the DNA whereby 30 µl of nuclease-free water were added and incubated at room temperature for 5 minutes. The column was centrifuged at full speed for 1 minute. DNA was stored long-term at 4°C in a microfuge tube.

2.3.2 Guthrie Card

DNA was extracted from archival dried blood spots using the QIamp Mini kit (Qiagen, Germany) using the manufacturer's instructions with some modifications. Briefly, 10 circles of blood-stained Guthrie Card (GC) spots were punched using an ethanol-sterilised stainless-steel hole puncher. PBS (180 µl) and Protease (20 µl) (Qiagen,

Germany) were added to the GC punches in a 1.5 ml microfuge tube and mixed thoroughly by vortexing. The tube was mixed on a plate shaker for 20 minutes and incubated at 56°C overnight. After the overnight incubation, the tube was homogenised by a TissueLyser II (Qiagen, Germany) for 30 seconds at 25 l/s to lyse the GC spots with the tungsten bead. The tube was centrifuged at maximum speed for 1 minute to pellet the homogenised GC. The tungsten bead was removed carefully from the microfuge tube and the supernatant was separated from the mulched GC and collected into a clean microfuge tube. To the supernatant, 200 mL of AL buffer were added, and the mixture was incubated at 65°C for 20 minutes. To the mixture, 200 µl of 100% ethanol were added followed by a second incubation at room temperature for 30 minutes. The entire volume was transferred to a QIamp mini column (Qiagen, Germany) and centrifuged at 6000 rcf for 30 seconds. The flow-through was discarded and two washes were performed using 500 µl of AW1 and AW2 buffer. The column was centrifuged at 13,000 rcf for 5 minutes to remove any residual wash buffer and transferred to a clean microfuge tube. A double elution was performed whereby, 100 µl of nuclease-free water were added to the column and left for incubation at room temperature for 2-5 minutes and centrifuged at 6000 rcf for 3 minutes to elute the DNA. The elution step was repeated with 50 µl of nuclease-free water to obtain a final elution volume of 150 µl. The DNA was stored long-term at 4°C.

2.4 DNA quantification using Qubit Assay

DNA was quantified using the Qubit dsDNA broad-range (BR) assay (Thermo Fisher Scientific, USA) using the standard protocol. Prior to DNA quantification, all solutions were equilibrated to room temperature. A Qubit working solution was prepared using Qubit dsDNA BR reagent and Qubit dsDNA BR buffer (1:200). For each standard, 190 µl of the working solution were added to 10 µl of the standard DNA. For each sample,

199 µl of the working solution were added to 1µl of sample DNA. The microcentrifuge tubes were mixed by vortexing for 10-15 seconds, spun down and incubated for 2 minutes at room temperature. The fluorometer was calibrated using standard 1 (0 ng/µl) and standard 2 (100 ng/µl). For each DNA sample, two replicate readings were taken using a fluorometer, and an average DNA concentration was calculated.

2.5 Genome-wide DNA methylation profiling

Genome-wide DNA methylation was measured using the Infinium HumanMethylation450K (HM450K) BeadChip assay as per the manufacturer's instructions (Illumina, USA). Sample preparation for the HM450K assay was done following an in-house developed workflow (Wong *et al*., 2015) that included a novel quality control (QC) checkpoint as shown in Figure 2.1.

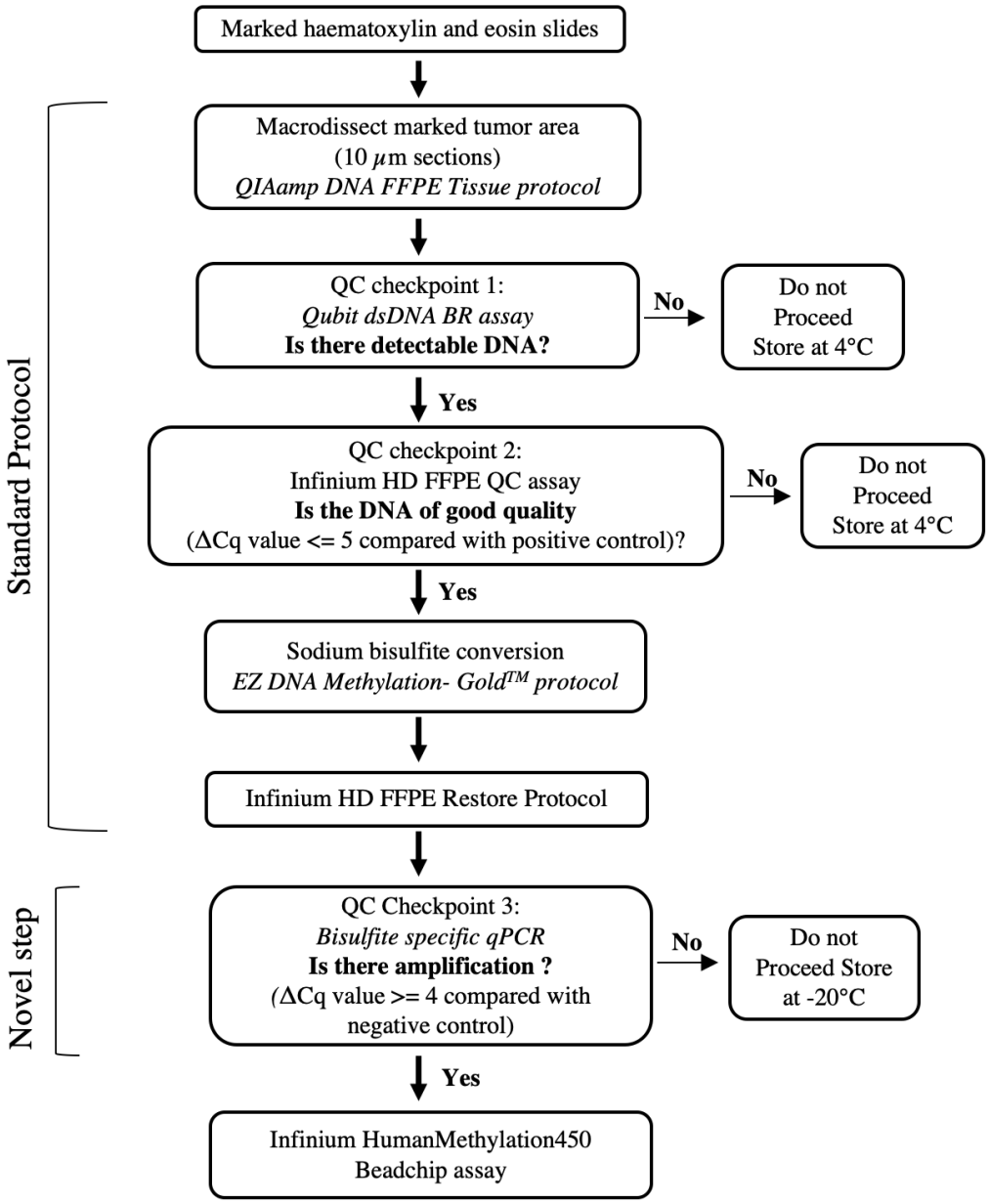


Figure 2.1: Sample preparation workflow for the Illumina HumanMethylation 450K BeadChip assay, adapted from (Wong *et al.*, 2015).

Flow chart illustrating the workflow for measuring the genome-wide tumour DNA methylation from formalin-fixed paraffin-embedded (FFPE) tumour derived DNA. Tumour DNA was first prepared by macrodissecting the tumour enriched area on the FFPE slide. The DNA was quantified using the Qubit dsDNA BR assay (QC checkpoint 1) and the DNA quality was estimated using the Infinium HD FFPE QC qPCR assay (QC checkpoint 2). Tumour samples that passed the initial quality checks were proceeded to sodium bisulfite conversion and restoration. The restored DNA samples were assessed on a final QC, (Bisulfite specific qPCR) developed in-house (QC checkpoint 3). The samples that passed the final QC were proceeded to the HM450K assay.

2.5.1 Infinium HD FFPE QC qPCR assay

To assess the quality of the DNA samples, a qPCR-based assay was performed prior to the HM450K methylation assay using Infinium HD FFPE QC Kit (Illumina, USA) according to the manufacturer's instructions. Briefly, the DNA was diluted to 1ng/µl and a 100-folds dilution of the QC Template was prepared using nuclease-free water. PCR reaction (10 µl) was prepared using the qPCR master mix and QC Primers. The following qPCR cycling program was used: initial incubation at 50°C for 2 minutes, enzyme activation at 95°C for 10 minutes, followed by 40 cycles of DNA denaturation at 95°C for 30 seconds, primer annealing at 57°C for 30 seconds and extension at 72°C for 90 seconds. All the samples were assayed in triplicate. The Cq values (number of quantification cycles) were obtained for each well and average Cq values were calculated for the DNA sample and the QC Template. ΔCq was calculated for all the sample DNA by subtracting the average Cq value for the QC Template from the average Cq value for the sample. Samples with ΔCq value ≤ 5 were proceeded to the next step.

2.5.2 Sodium bisulfite conversion

To differentiate the methylated cytosine from unmethylated cytosines, the tumour DNA was bisulfite converted prior to DNA methylation profiling. This process leads to the deamination of unmethylated cytosines to uracils, whereas the methylated cytosines remain unchanged (Figure 2.2). In subsequent PCR reaction, the uracils are amplified as thymines, whereas methylated cytosines get amplified as cytosines, thus the methylated cytosines are differentiated from the unmethylated cytosines.

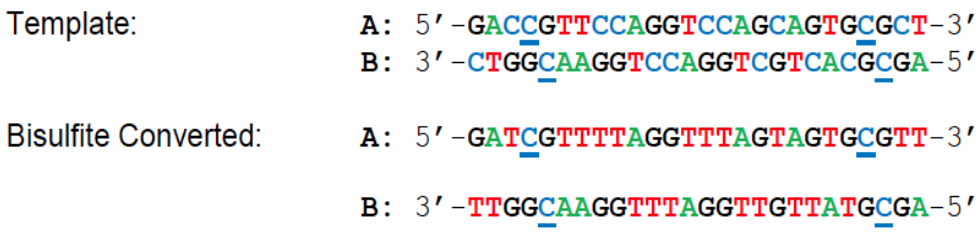


Figure 2.2: Sodium bisulfite conversion (Zymo Research, USA)

Graphics illustrating the process of sodium bisulfite conversion. "Template" in the diagram represents a DNA sequence before sodium bisulfite conversion and "Bisulfite Converted" represents the DNA sequence after sodium bisulfite conversion. A, T, G and C in the Template and Bisulfite Converted DNA represent the nucleotides adenine, thymine, guanosine and cytosine, respectively. Methylated cytosines are represented as "C" and unmethylated cytosines are represented as "C". After bisulfite conversion of the template DNA, methylated cytosines remain as it is, whereas the unmethylated cytosines convert to thymine, thus differentiating the unmethylated cytosines from methylated cytosines.

Up to 500 ng of tumour DNA was bisulfite converted using Zymo Gold EZ-DNA kit (Zymo Research, USA) following the manufacturer's instructions. Briefly, 130 µl of CT Conversion Reagent were added to 20 µl of tumour DNA and mixed by pipetting. The mixture was incubated using the following protocol in a thermal cycler: 98°C for 10 minutes, 64°C for 2.5 hours and a hold step at 4°C (for up to 20 hours). Following incubation, the total volume was transferred to a Zymo-Spin™ IC Column, placed into a collection tube containing 600 µl of M-binding buffer and mixed properly by inverting. The tube was centrifuged at full speed for 30 seconds and the flow through discarded. The Zymo-Spin™ IC Column was washed with 100 µl of M-wash buffer. To the column, 200 µl of M-Desulphonation buffer were added and left to incubate at room temperature for 15-20 minutes. The column was washed twice using 200 µl of M-wash buffer and placed into a clean 1.5 ml microfuge tube. The bisulfite-converted DNA was eluted in M-Elution Buffer. The bisulfite converted DNA was stored at -20 °C.

2.5.3 DNA restoration

DNA restoration is carried out to repair the degraded FFPE DNA. For DNA restoration, the Infinium HD FFPE Restore kit (Illumina, USA) was used and the restoration was performed according to the manufacturer's instructions. Briefly, to 8 µl of bisulfite converted tumour DNA, 4 mL of freshly prepared 0.1N NaOH were added and incubated at room temperature for 10 minutes. After incubation, 34 µl of Primer Pre-Restore Reagent and 38 µl of Amp Mix Restore Reagent were added to the mixture and mixed thoroughly by inversion followed by a centrifuge at 280 rcf for 1 minute and an incubation at 37°C for 1 hour. After incubation, the mixture was centrifuged at 280 rcf for 1 minute and 7 volumes (560 µl) of Zymo DNA Binding Buffer were added to each volume of DNA and mixed by pipetting. The mixture was transferred to a Zymo-Spin™ I-96 Plate (Zymo Purification Kit) mounted on a collection plate and centrifuged at 2250 rcf for 2 minutes and the flow through was discarded from the collection plate. The Zymo-Spin I-96 plate was mounted on a new collection plate and to the DNA sample well, 600 mL of Zymo Wash Buffer (with 100% ethanol added) was dispensed. The plate was centrifuged at 2250 rcf for 2 minutes and the flow through was discarded from the collection plate. A clean 0.8 ml 96-well plate was prepared with the Zymo-Spin I-96 plate mounted on it. To each sample well, 13 µl of Elution Restore Buffer Reagent were dispensed directly and the plate was incubated at room temperature for 5 minutes. The sample plate assembly was centrifuged at 2250 rcf for 1 minute to elute the DNA. The plate containing approximately 10 µl of eluted DNA was sealed with an adhesive foil seal and incubated for 2 minutes at 95°C on a heat block. Immediately after incubation, the sample plate was transferred to an ice bucket and incubated for 5 minutes (making sure that the bottom of the wells was in contact with ice). Keeping the sample plate on ice, 10 µl of Convert Master Mix Reagent were added to each sample well and the plate was

vortexed for 1 minute at 1600 rpm and centrifuged at 280 rcf for 1 minute followed by an incubation at 37°C for 1 hour. After incubation, the plate was centrifuged at 280 rcf for 1 minute and 7 volumes (140 µl) of Zymo DNA Binding Buffer were added to each volume of DNA in the sample plate. The sample mixture was mixed by pipetting and then transferred to another Zymo-Spin I-96 Plate mounted on a collection Plate. The Zymo-Spin I-96 Plate was centrifuged at 2250 rcf for 2 minutes and the flow through in the collection plate was discarded. To the sample well of the Zymo-Spin I-96 Plate, 600 µl of Zymo Wash Buffer (with 100% ethanol added) were added and the assembly was again centrifuged at 2250 rcf for 2 minutes and the flow through discarded. A new 0.8 ml 96-well microtiter plate was prepared, and the Zymo-Spin I-96 Plate was mounted on the sample plate. To each well of the Zymo-Spin I-96 Plate column matrix, 12 mL of nuclease-free water was dispensed directly, and the plate was incubated at room temperature for 5 minutes. After incubation, the plate assembly was centrifuged at 2250 rcf for 1 minute to elute the DNA.

2.5.4 Bisulfite specific qPCR

After bisulfite conversion and restoration, the samples underwent a final quality check before the methylation assay. The QC was performed using bisulfite-specific PCR assay designed in-house (Wong *et al.*, 2015).

For the qPCR assay, bisulfite converted DNA specific Forward Primer: 5′ tAA GGT AtA ATt AGA GGA TGG GAG GGA t and Reverse Primer: 5′ aaC AAA CTC Aaa TAa AAT TCT TCC TC were designed (Wong *et al.*, 2015). Lower-case letters in the primer sequences correspond to bisulfite converted cytosines. The primers were

designed to amplify a 134 bp region (hg19: chr17:41,277,493-41,277,626) within the promoter of the *BRCA1* gene (GenBank: L78833.1).

For the assay, a 10 µl reaction was prepared using 5 µl FastStart SYBR Green dye (Roche Diagnostics, Switzerland), 0.3 µl each of forward and reverse primers (10µM) (Integrated DNA technologies, USA) and 3 µl of diluted restored bisulfite converted DNA (1:6 in nuclease-free water). The reaction volume was equilibrated to 10 µl with nuclease-free water. A non-bisulfite converted; high molecular weight genomic DNA isolated from the U266 multiple myeloma cell line sourced from the Peter MacCallum Cancer Centre was used as a negative control. The following qPCR cycling conditions were used for the assay: initial polymerase activation at 95°C for 5 minutes followed by 40 cycles of DNA denaturation at 95°C for 10 seconds, primer annealing at 60°C for 30 seconds and extension at 72°C for 90 seconds. All the samples and controls were assayed in duplicate and the average Cq value (number of quantification cycle) was obtained for each sample. To determine the presence and bisulfite-conversion efficiency, ΔCq was calculated by subtracting the average Cq value for each sample DNA from the average Cq value of the negative control. Restored and bisulfite-converted tumour-derived DNA samples with a ΔCq value of ≥ 4 were progressed to the Infinium HM450K assay.

2.5.5 Loading samples on the HM450K BeadChip

The HM450K assay design included the following controls: i) the multiple myeloma cell line (U266) DNA that was included as an internal control and ii) one technical replicate from restored bisulfite converted FFPE tumour-enriched DNA to test for reproducibility. Each technical replicate was placed on a different BeadChip to test for possible effects on data generated on two different BeadChips.

Before loading the samples on the HM450K array, 4 µl of bisulfite converted and restored DNA sample were denatured by mixing with freshly prepared 0.1N NaOH and neutralized using the Random Primer Mix followed by an incubation for 20-24 hours at 37°C where the denatured DNA was isothermally amplified. After 24 hours incubation, the amplified DNA was fragmented using a controlled enzymatic process. The fragmented DNA was precipitated using 100% 2-propanol and centrifuged at 3000 rcf at 4°C for 20 minutes to collect the DNA. After centrifugation, the supernatant was discarded by inversion (taking care to not disturb the DNA pellet) and left uncovered and inverted on a tube rack for 1 hour at room temperature to air-dry the DNA pellet.

The precipitated DNA was resuspended in 23 µl RA1 and incubated for 1 hour at 48°C. The resuspended DNA was incubated at 95°C for 20 minutes to denature the DNA and then left at room temperature for 30 minutes to cool down. The BeadChips (Illumina, USA) were removed from 4°C storage and hybridisation chambers (Illumina, USA) were prepared as per manufacturer's instructions. The BeadChip was placed in the hybridisation chamber insert containing 400 µl of humidifying buffer, PB2. Prior to loading the BeadChip, a sample sheet was prepared and BeadChip barcode and chip position was recorded for each sample. Using a multichannel pipette, 15 µl of each DNA sample were dispensed in the sample inlet on the BeadChip at assigned positions. The hybridisation chamber inserts containing the BeadChip was placed in the Illumina hybridisation chamber and incubated at 48°C in the Illumina Hybridisation Oven (with the rocker function on) for 20-24 hours. After incubation, the BeadChip was washed with PB1 to remove any unhybridized and non-specifically hybridised DNA. The flow-through chamber assembly was prepared and proceeded to the extension and staining step which was performed on the TECAN liquid handler (Tecan, Switzerland). Post-staining

and extension, the BeadChip was washed with 310 ml of PB1 and 310 ml of XC4 and dried under vacuum (675 mm Hg) for 50-55 minutes. The underside of the beadchip was wiped to remove any excess XC4 and scanned on Illumina iScan.

2.6 Methylation data pre-processing and normalisation

Raw intensity files (.IDAT files) for all the samples were imported in the R programming software (version 3.6.1) using the Bioconductor package *minfi* (Aryee *et al.*, 2014) and a standard workflow was followed for pre-processing and normalising the data (Figure 2.3).

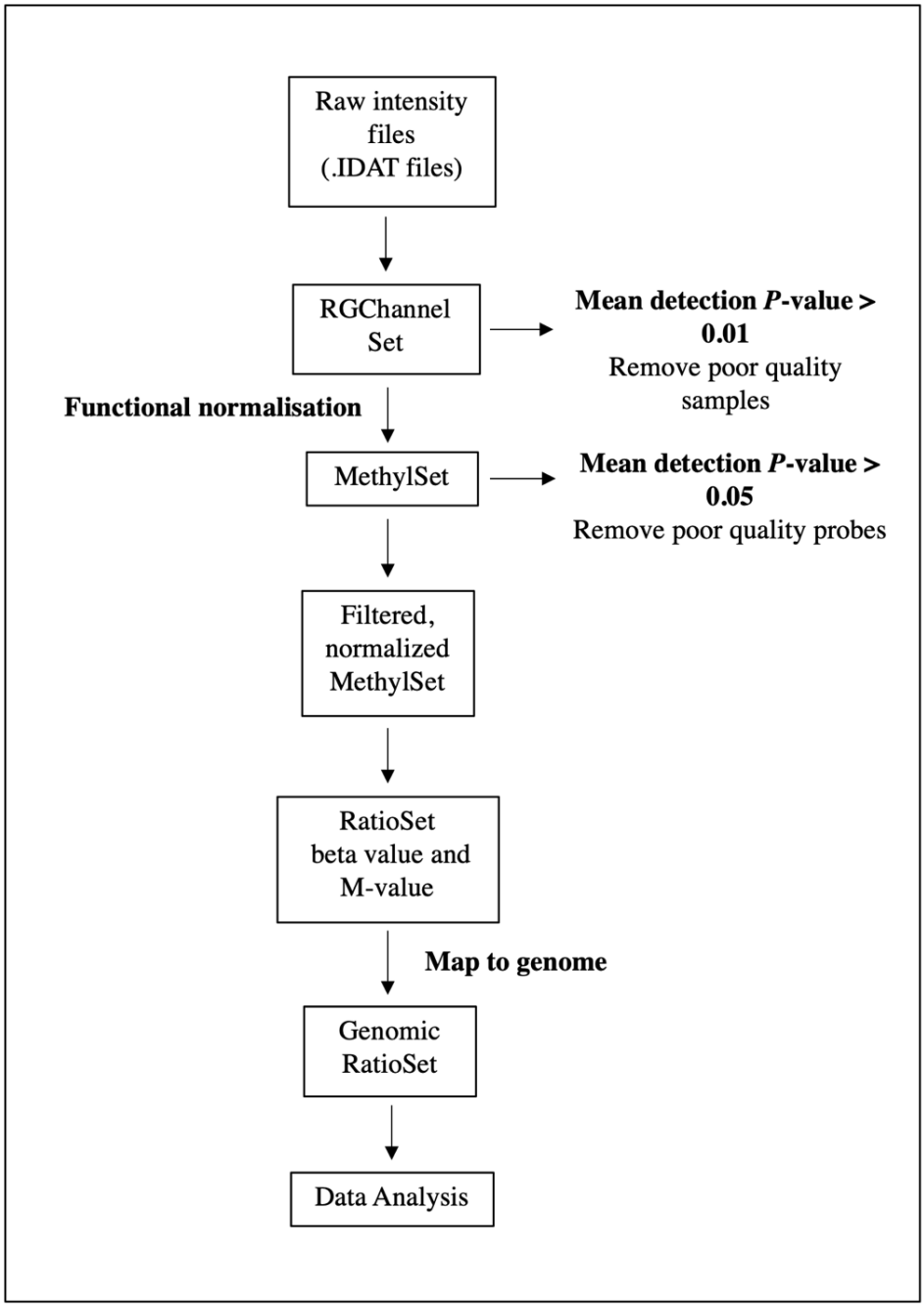


Figure 2.3: DNA methylation data pre-processing and normalisation workflow.

Flowchart illustrating the DNA methylation data pre-processing and normalisation workflow. Intensity signals from both red and green channels were read from the raw intensity files (.IDAT files) and stored in a RGChannelSet object. Detection *P*-values were obtained, and poor-quality samples (detection *P*-value > 0.01) were removed from further analysis. The filtered RGChannelSet object was then normalised using the functional normalisation method. The normalised data stored in the MethylSet object was checked for poor quality probes and the probes with mean detection *P*-value > 0.05 were removed from further analysis. Filtered and normalised data were used to calculate the RatioSets (M values and beta-values) for the samples. The M values and beta-values were annotated to the hg19 human genome and were used for further analyses.

The data quality was first evaluated by assessing the detection *P*-value. Detection *P*-value is an indicative factor for the quality of the signals and in *minfi*, it is calculated by comparing the total signal (Methylated+Unmethylated) for each probe to the background signal level (estimated from the negative control probes). Samples with a mean detection *P*-value of more than 0.01 were deemed poor quality and were removed from further analysis.

Functional normalisation (FNORM) method was used to normalise the methylation data. This normalisation method corrects for both within array (technical bias between type I and type II probes) as well as between array unwanted variations (Fortin *et al.*, 2014). It also applies a background adjustment method, "noob" that corrects for any dye-bias (Triche *et al.*, 2013). FNORM also corrects for potential batch effect and thus no further batch correction was performed on the data. After normalisation, CpG probes with a mean detection *P*-value of more than 0.05 in one or more samples were considered unreliable and were removed from further analyses. No further filtering was performed on the data. M-values and beta-values were calculated from the normalised and filtered data. For all statistical analyses, M-value was used and the beta-value was mainly used for data exploration and visualisation as suggested in (P Du *et al.*, 2010). Methylation level (beta-value) of more than 0.50 was defined as hypermethylated and beta-value of less than 0.50 was defined as hypomethylated.

2.7 Tumour purity estimation

Tumour purity was estimated using the *R* tool *InfiniumPurify* (Qin *et al.,* 2018) that takes methylation beta-values of the tumour samples and uses the methylation levels of pre-selected informative differentially methylated CpG sites (iDMCs) identified from

TCGA data (when normal data is not available) to estimate tumour purity for each tumour sample by density evaluation of Gaussian kernel. Tumour purity estimate was obtained as the proportion of tumour cells in each sample.

2.8 The Cancer Genome Atlas data

Raw DNA methylation data (.IDAT files) for 659 breast cancer cases (168 ILBC and 491 IDBC), were downloaded from TCGA legacy database (Study Accession: phs000178) using the R package *TCGABiolink* (Colaprico *et al.*, 2016). The methylation data was pre-processed and normalised similarly as the study set and methylation values (beta-values and M-values) were obtained for all cases at 440,380 CpG positions across the genome. Survival data was retrieved for 159/168 (95%) ILBC cases. Gene expression data in the form of normalised counts (RNA sequencing-Illumina Hi-Seq) was retrieved for 159/168 (95%) ILBC cases.

2.9 Statistical analyses

2.9.1 Differential methylation analysis

Differentially methylated positions (DMPs) between the comparison groups were identified by a probe-wise differential methylation analysis using the *Limma* Bioconductor package (Smyth, 2005). A linear regression model was implemented on the M value matrix of the samples to obtain moderated t-statistics and associated *P*-values for each CpG position. The *P*-values were corrected for multiple testing using Bonferroni correction and a false-discovery rate (fdr) cut-off of 0.01 (Bonferroni, 1935).

Differentially methylated regions (DMRs) were identified using the *DMRcate* package in R (Peters *et al.*, 2015). To calculate the DMRs, a gaussian smoothing estimate was applied to group the adjacent DMPs within 1000 bp window. The smoothed test statistics was modelled using the Satterthwaite method and *P*-values were computed based on this model (Satterthwaite, 1946). The *P*-values were corrected using a fdr cut-off of 0.01. The consecutive significant CpG sites (1000 bp from each other) were agglomerated together and the regions were defined as DMRs.

2.9.2 Variable methylation analysis

Variable methylation analysis was performed using the *DMRcate* (Peters *et al.*, 2015) package in R. To identify the variably methylated regions (VMRs), the variance of M values was computed across the samples and gaussian smoothing was applied to the resulting per-CpG-site test statistics using the default *DMRcate* options. *DMRcate* uses the method of Satterthwaite to smooth test statistics and derive respective *P*-values (Satterthwaite, 1946). Nearby significant CpG sites were collapsed in clusters using a bandwidth of 1000 bp. The clusters that showed the highest variability in DNA methylation (i.e., regions with a minimum adjusted *P*-value (minfdr) of less than 10^{-8}) were defined as VMRs.

2.9.3 Gene set enrichment analysis

Gene set enrichment analysis was performed using a web-based tool *Metaspace* using the default settings (Yingyao Zhou *et al.*, 2019). Pathway and gene set enrichment analysis were carried out using the following ontology sources: Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathway (Kanehisa *et al.*, 2017), Gene Ontology (Ashburner *et al.*, 2000) and Reactome Gene Sets (Croft *et al.*, 2010). All genes in the genome were used as the enrichment background. Pathways and biological terms with a

P-value < 0.01, a minimum count of 3 genes, and an enrichment factor > 1.5 (the ratio between the observed counts and the counts expected by chance) were collected and grouped into clusters.

2.9.4 Survival analysis

Survival analysis was performed using the *Survival* package in R (Therneau, 2014). Cox proportional hazards regression models were used to calculate the HRs and 95% CI for the association between DNA methylation levels (M values) and risk of death. Survival curve were constructed using the Kaplan-Meier estimator, and survival difference were compared using the log-rank test. Information on death in the MCCS cohort, was collected from the following sources: Victorian Birth, Death and Marriages (Birth Deaths and Marriages Victoria, 2020) and notified death recorded in Victorian Cancer Registry (Victorian Cancer Registry, 2020). For kConFab and ABCFR, information on death data collection and linkage was not available. For the MCCS that made up to ~90% of the sample size in this study, the latest linkage was done on 31 March 2017 and was considered to be complete up to 31 December 2016. Overall survival was defined as the time (in years) elapsed between breast cancer diagnosis and death (from breast cancer or any other cause) or end of follow-up. Follow-up started at the date of diagnosis and ended at the date of death or end of follow-up (31 December 2016), whichever came first.

2.9.5 Unsupervised cluster analysis

Unsupervised cluster analysis was performed based on the methylation levels (M values) of the samples across 449,005 CpG positions and the Euclidean distance was estimated using *ward.D2* minimum variance method (Murtagh & Legendre, 2014). The samples were grouped based on the similarity in their methylation levels. In this method,

the clustering begins with all samples as individual clusters. A dissimilarity matrix is formed which is the squared Euclidean distance between the cluster means. The clusters are repeatedly merged into a pair of clusters such that when merged; there is a minimum increase in total within-cluster variance (bottom up). The merging of clusters continues until a single group including all samples (the top of the tree) is defined.

2.9.6 Statistical tests for testing associations

Pearson's chi-square tests the association between categorical variable, whereas t.test and ANOVA were used to test the association with continuous variables. A *P*-value of < 0.05 was considered significant.

2.10 Whole-exome sequencing

Whole-exome sequencing (WES) was performed for somatic mutation profiling of a subset of selected ILBC cases. Libraries for WES were prepared using SureSelect XT low input library preparation kit (Agilent, USA) as per manufacturer's instructions and SureSelect clinical research exome v2 (CRE v2) (Agilent, USA) was used as the target capture.

2.10.1 NGS FFPE QC qPCR

Prior to library preparation, the quality and amplifiable DNA quantity of tumour DNA and germline DNA samples were assessed using the NGS FFPE qPCR QC assay (Agilent, USA), as per manufacturer's instructions. The workflow for this assay is summarised in Figure 2.4.

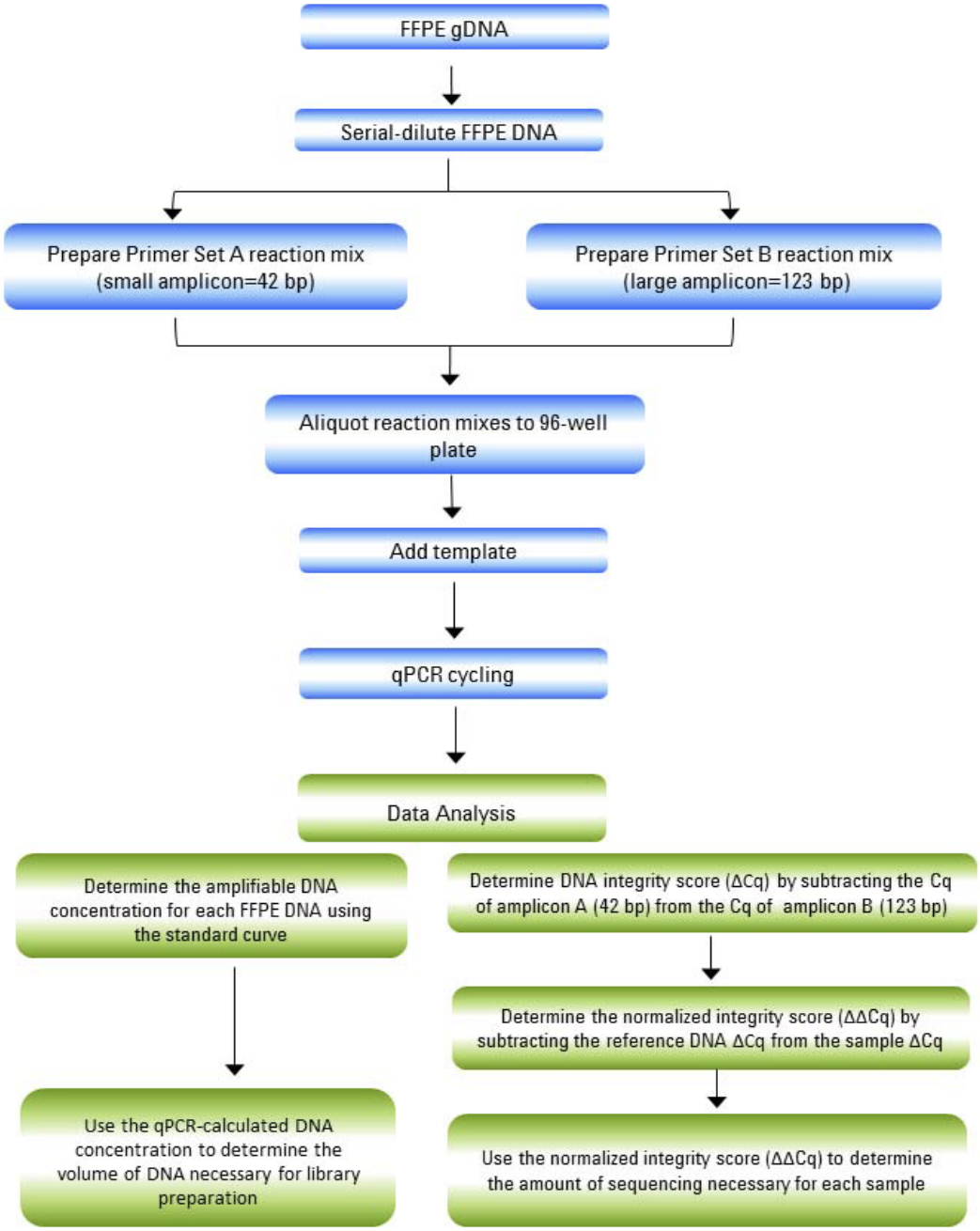


Figure 2.4: NGS FFPE QC qPCR assay workflow (Agilent, USA).

Flowchart illustrating the NGS FFPE qPCR QC assay and the data analysis workflow. The DNA samples to be assessed were serial diluted to 125 pg/µl. The diluted samples were then amplified with Primer set A and Primer set B. Standard curve analysis was used to calculate the amplifiable DNA quantity and the DNA quality was determined by comparing the amplification of each sample to the reference DNA.

Briefly, the initial DNA concentration was estimated for each sample using the Qubit dsBR assay kit (Thermo Fisher Scientific, USA) (as described in section 2.4). The DNA sample was serial diluted to 125 pg/µl as instructed in the protocol and the dilution factor (final volume/sample volume) was recorded for each sample. For the qPCR reaction, two sets of reaction mix were prepared, one for Primer Set A that targets a single-copy DNA region of the human genome and produces a 42 bp amplicon and another for Primer Set B that targets the same DNA region but produces a 123 bp amplicon. Each qPCR reaction (for 1 sample) contained 10 µl of 2× Brilliant III SYBR Green qPCR Master Mix, 1 µl of Primer Set A or Primer Set B, 0.3 µl diluted reference dye (freshly prepared, 1:500 dilution) and 4 µl of DNA sample. The volume was equilibrated to 20 µl using the nuclease-free water.

The reaction plate was set in such a way that the pre-diluted DNA standards: DNA standard 1 (2500 pg/µl); DNA standard 2 (625 pg/µl); DNA standard 3 (156.25 pg/µl); DNA standard 4 (39.06 pg/µl); and DNA standard 5 (9.77 pg/µl) were only amplified with Primer Set A to generate a standard curve for each qPCR run. Sample DNA and a positive control DNA were amplified using both Primer Set A and Primer Set B. All the reactions were set up in triplicate. QuantStudio 7 Flex qPCR system (Thermo Fisher Scientific, USA) was used for the assay and the run was set for real-time detection of SYBR Green fluorescence at the annealing and extension steps, reporting of quantification cycle (Cq), and subsequent standard curve analysis. The qPCR thermal cycling profile was as follows: 95°C for 3 minutes and 40 cycles of DNA denaturation at 95°C for 10 seconds and annealing and extension at 63°C for 20 seconds.

The quantity of amplifiable DNA present in a DNA sample was determined by the standard curve analysis. A standard curve was generated (plot of initial DNA quantity versus Cq), using the DNA standards amplified using Primer Set A (Figure 2.5). It was verified that the standard curve has an R^2 value > 0.98 and an amplification efficiency between 85% and 110%. The concentration of amplifiable DNA was calculated by dividing the amplifiable DNA quantity by 4 µl (volume of sample DNA added to each qPCR reaction) and then multiplied by the dilution factor recorded for each sample.

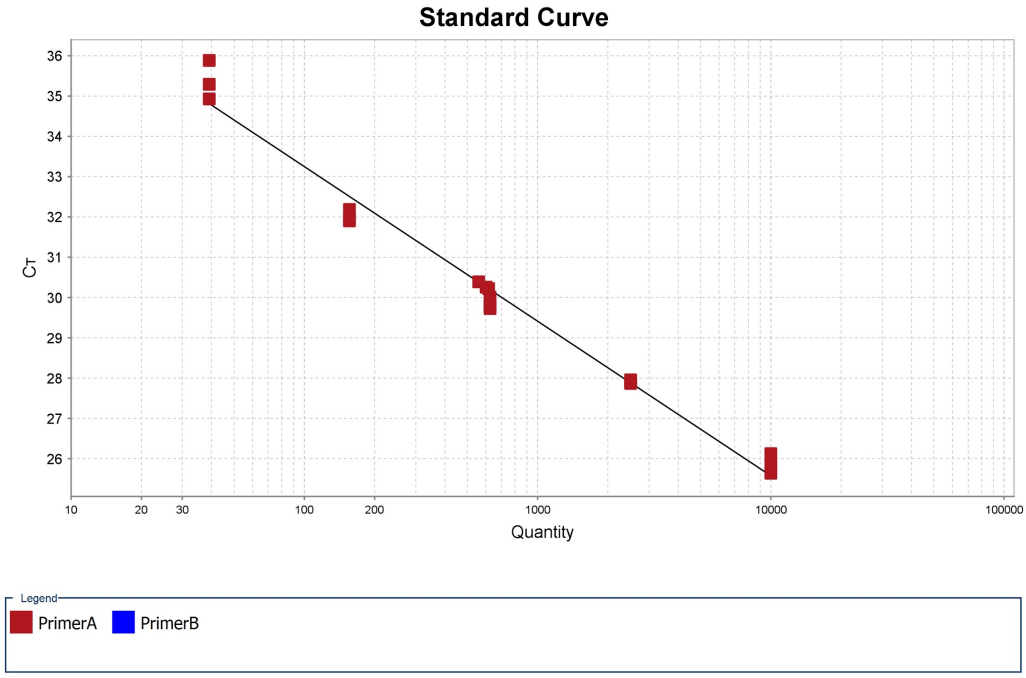


Figure 2.5: Standard curve.

Standard curve showing five points of a 4-fold dilution series of DNA standards (1-5), performed in triplicate, amplified using the Primer Set A. The standard curve was generated by plotting the threshold cycle (C_T) values (on the y-axis) against relative input standard DNA quantities (on the x-axis).

The DNA quality was estimated by assessing the relative amplification of the sample DNA using Primer Set A, compared with Primer Set B and the amplification of the reference DNA by Primer Set A and Primer Set B. For each sample and the reference DNA, integrity score or ΔCq was calculated by subtracting the Cq values from reaction B from the Cq values from reaction A for each sample. A normalised DNA integrity score or ΔΔCq was calculated for each DNA sample by subtracting the reference DNA ΔCq from the sample ΔCq as, ΔΔCq = ΔCqSample - ΔCqRef.

2.10.2 Shearing the DNA using Covaris

Before library preparation, tumour DNA and germline DNA were sheared to a target fragment size of 150-200 bp by mechanical shearing using the Covaris sonicator (Covaris, USA). For this, 200 ng of sample DNA was normalised to a final volume of 50 µl in 1X Low TE Buffer. The Covaris tank was filled with MilliQ water to the appropriate level and the degas process was initiated in the SonoLAB software (Covaris, USA). The chiller temperature was set to 2-5°C. The instrument was left to degas for at least 30 minutes before use. After 30 minutes, 50 µl of sample DNA was loaded in the Covaris microtube (130 µl, pre-slit snap cap) using a normal pipette tip through the pre-split septa of the cap. It was ensured that no bubbles were introduced into the bottom of the tube. The microtube was secured in the Covaris tube holder and the FFPE and genomic DNA were sheared using settings provided in protocol with different treatment times for FFPE (240 seconds) and genomic DNA (2x 120 seconds). Once the shearing was complete, the sheared DNA from the microtube was transferred to a fresh 1.5 ml microfuge tube and kept on ice.

2.10.3 Library Preparation

Library preparation for WES was performed using the using the SureSelect XT Low Input library preparation kit (Agilent, USA) as per the manufacturer's instructions. SureSelect CRE v2 (Agilent, USA) was used as the target capture.

Input DNA quantity for library preparation was determined based on the DNA integrity score, ΔΔCq as determined in the NGS FFPE qPCR QC assay (section 2.10.1). For samples with $\Delta\Delta Cq \leq 1$, the Qubit-based quantity estimate was used and for the samples with $\Delta\Delta Cq \geq 1$, qPCR-based quantity estimate was used to determine the input DNA quantity.

The library preparation workflow is illustrated in Figure 2.6. Briefly, to the sheared DNA, dA-tail and end-repairing were performed. A unique molecular barcode sequence that are short random nucleotides also called unique molecular index (UMI) was ligated into each DNA fragment and were used to mitigate the PCR errors in this study. The adaptor ligated library was purified using 80 µl of homogeneous AMPure XP beads. For this, AMPure beads and the library were mixed well by pipetting and incubated at room temperature for 5 minutes. The mixture was kept on a magnetic separator device and left for 5-10 minutes for the solution to clear. Keeping the mixture plate on the magnetic stand and without touching the beads, the cleared solution was removed and discarded. To each well, 200 µl of freshly prepared 70% ethanol were dispensed, and any disturbed beads were allowed to settle for 1 minute. The ethanol was removed and discarded. The ethanol wash was repeated, and the plate was left to air dry to remove any residual ethanol.

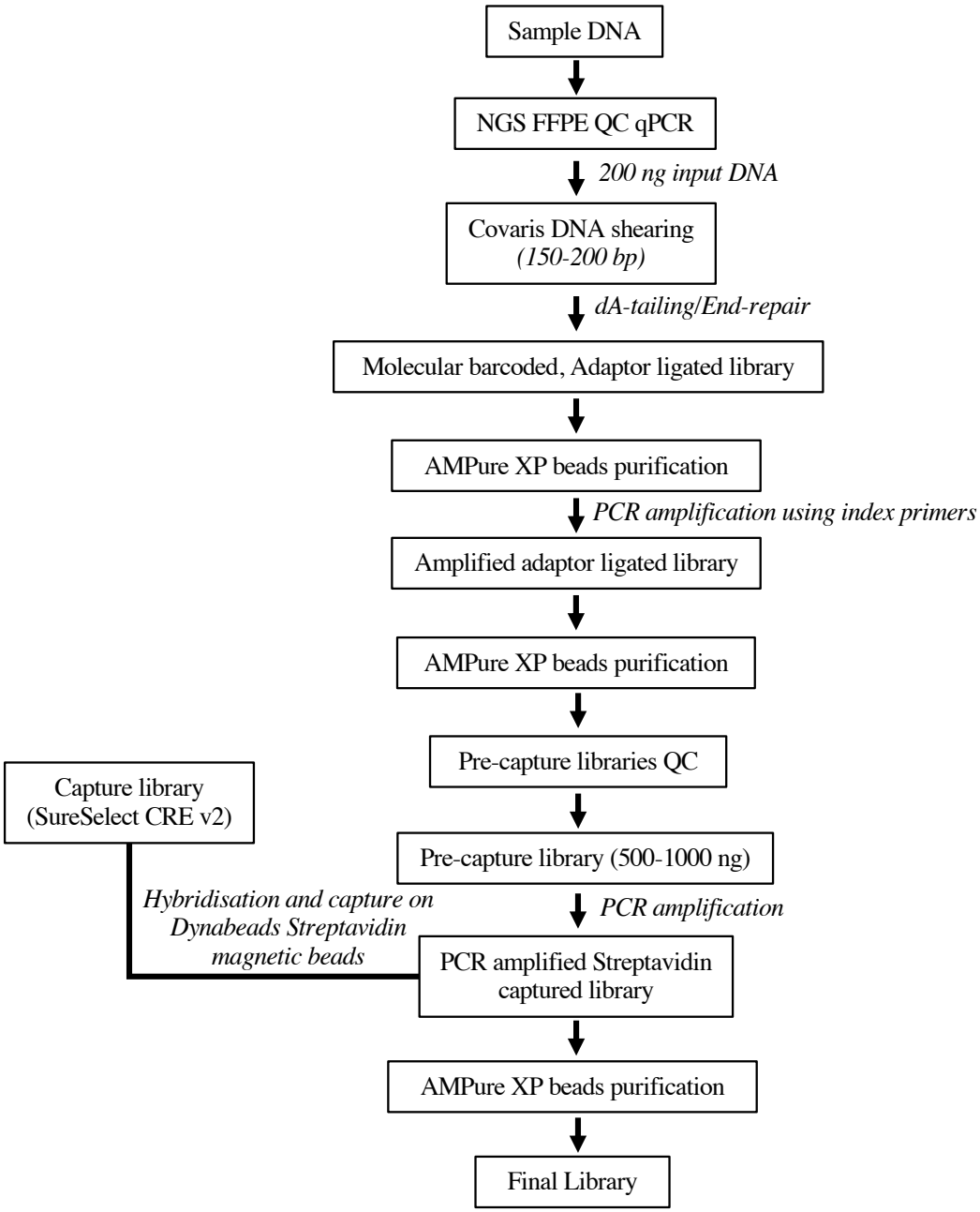


Figure 2.6: Library preparation workflow.

The flow diagram illustrating the whole-exome sequencing (WES) library preparation workflow. The input DNA quality and quantity were estimated using NGS qPCR QC assay (Agilent, USA). The DNA was fragmented using Covaris mechanical shearing and sequencing adaptors were ligated to each library in the dA tailing and adaptor ligation step. The adaptor ligated library was purified using the AMPure XP beads and was PCR amplified. The quality and quantity of the pre-capture library was assessed using the TapeStation and D1000 ScreenTape. Hybridisation reaction was performed, and the captured library was PCR amplified. The captured and amplified library was purified using the AMPure XP beads to give the final sequencing library.

After the plate was dry, 35 µl of nuclease-free water were added, vortex mixed and incubated at room temperature for 2 minutes. The plate was transferred to the magnetic rack and left for 5 minutes or until the magnetic beads were completely bound to the magnet. The cleared supernatant (adaptor ligated library) was transferred to a fresh PCR plate and amplified using the SureSelect XT Low Input Index Primers. For FFPE DNA samples, 11 PCR cycles were performed, whereas for all high molecular weight DNA samples, 8 PCR cycles were performed. Unique indexing primers were ligated to the library at this stage.

Once the PCR was complete, the library was purified using 50 µl of AMPure XP beads following the steps mentioned earlier. After the bead purification, the amplified adaptor ligated library was eluted in 15 µl of nuclease-free water. The quality and quantity of the pre-capture library was assessed on the 4200 TapeStation (Agilent, USA) using the D1000 ScreenTape (Agilent, USA). Library concentration was calculated by integrating the area under the peak as instructed in the TapeStation analysis protocol. The electropherogram for all the libraries were verified for the expected profile and DNA fragment peak. Library prepared from a high-quality DNA was expected to have a peak positioned between 300-400 bp, whereas for a low-quality DNA the peak was expected to range from 200-400 bp.

After the pre-capture library quality and quantity assessment, 500-1000 ng (calculated based on the TapeStation assessment) of pre-capture library was prepared in 12 µl nuclease-free water. Thermal cycler was programmed on the SureCycler (Agilent, USA) and the hybridisation reaction was performed as instructed in the protocol. The hybridisation mixture was transferred to a plate containing 200 µl of washed streptavidin

beads and mixed well by pipetting. The plate was incubated on a plate mixer at 1400-1800 rpm at room temperature for 30 minutes. After incubation, the plate was kept on the magnetic separator to collect the beads and the supernatant was removed and discarded. The beads were then resuspended in 200 µl of pre-warmed SureSelect Wash Buffer 2 and put on the magnetic separator to collect the beads. This wash was repeated and a total of six washes were performed. After the last wash, 25 µl of nuclease-free water were added to the library and the beads were resuspended by pipetting. The enriched DNA library was PCR amplified using the post-capture PCR thermal cycling program as instructed in the protocol and nine PCR cycles were performed as instructed in the protocol. Once the PCR amplification program was complete, the plate was kept on the magnetic stand and the supernatant (approximately 50 µl) was removed to a fresh 96-well PCR plate. The beads were discarded this time. The amplified captured library was purified using 50 µl of AMPure XP beads as described earlier. After purification, the final library was eluted in 25 µl of nuclease-free water. The quality and quantity of the final library were assessed on 4200 TapeStation (Agilent, USA) using high-sensitivity D1000 ScreenTape assay. Figure 2.7 shows a schematic diagram of the final sequencing library.

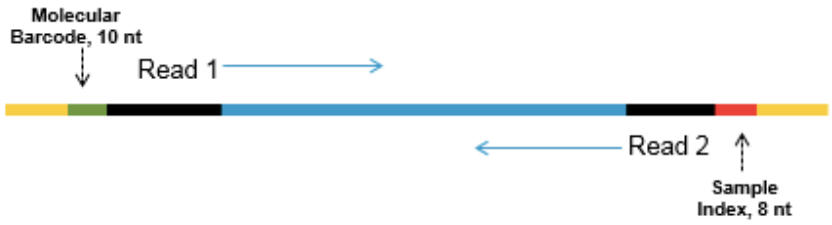


Figure 2.7: A schema showing the SureSelect XT Low Input sequencing library (Agilent, USA).

Content of SureSelect library sequence with sequencing adaptors and barcode ligated to 5' and 3' ends. Each fragment contains one target insert (blue) surrounded by the Illumina paired-end sequencing elements (black), the sample index (red), the molecular barcode (green) and the library bridge PCR primers (yellow).

2.10.4 Library pooling and sequencing

For multiplex sequencing, the FFPE DNA and germline DNA libraries were pooled in a of 4:1 ratio based on the data output requirement, which was a mean target depth of coverage of 150X and 50X for FFPE DNA and germline DNA samples, respectively. Starting with different concentrations, the following formula was used to determine the amount of each indexed library to be added in the pool.

$$\text{Volume of Index} = \text{V(f)} * \text{C(f)} / \text{number of libraries} * \text{C(i)}$$

where, V(f) is the final desired volume of the pool, C(f) is the final concentration (pg/µl) of all the libraries in the pool and C(i) is the initial concentration of each indexed library. The sequencing was performed on NovaSeq 6000 (Illumina, USA) on a single lane of S4 flow cell to generate 150 bp paired end reads.

2.11 Sequencing data processing and somatic variant calling

A WES somatic variant calling pipeline was implemented for processing the sequencing data and for identifying the somatic variants in tumour-normal analysis mode (Figure 2.8). Raw sequencing data was obtained as fastq files and an initial QC was performed using *FASTQC* (Andrews S, 2010) to generate a quality report that included basic metrics such as per base sequence quality, per base GC content and adapter content. After the initial QC, UMI tags were added using *fgbio* (https://github.com/fulcrumgenomics/fgbio) to get UMI annotated unmapped bam files. Multiple sequencing reads from PCR duplicates with identical UMI-tags were identified after read alignment and were collapsed together to generate a single consensus read thus, getting rid of the PCR duplicates. Adapter sequences were marked and removed and the

unmapped bam files were mapped to the reference genome (hg19) to get aligned bam files using *BWA mem* (H Li, 2013).

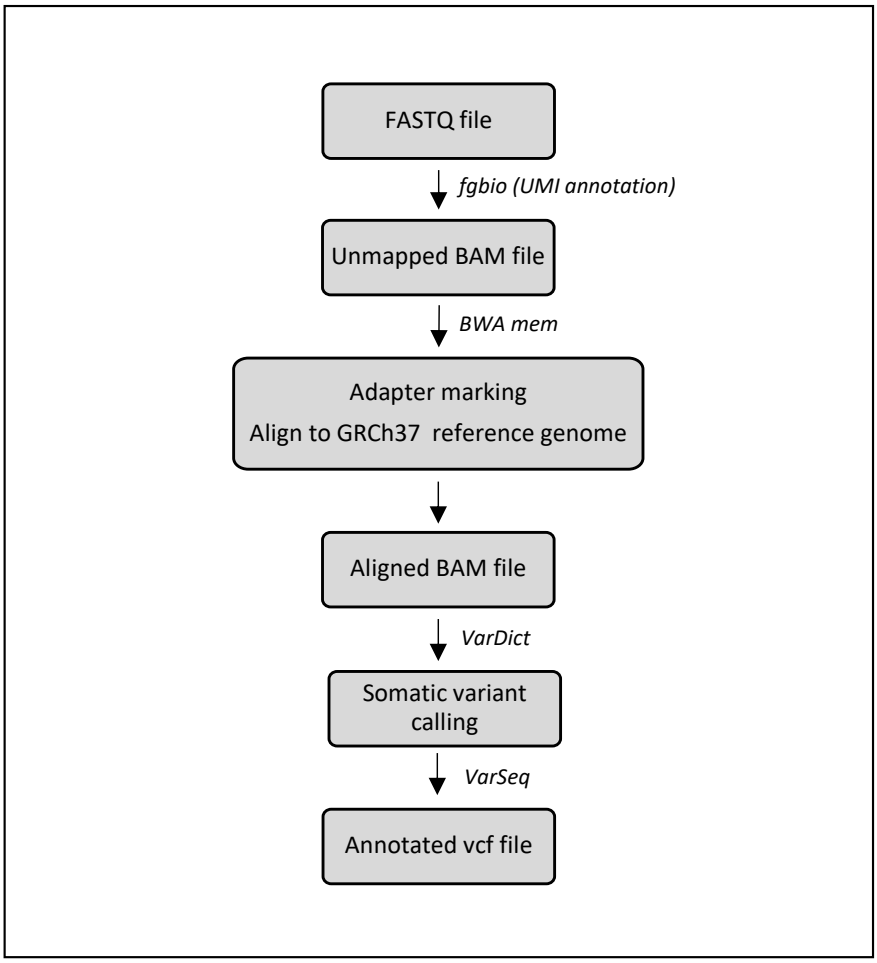


Figure 2.8: Whole-exome sequencing data processing and somatic variant calling. Flowchart illustrating the whole-exome sequencing and somatic variant calling workflow.

Somatic variant calling was performed using *VarDict* (Lai *et al.*, 2016) in paired tumour-normal analysis mode. *VarDict* default filters were used, which are as follows: Mean base quality > 22.5, Mean mapping quality > 0, Variant depth > 3, Total depth > 5, Allele frequency > 0.01 and *P*-value < 0.05. The variants satisfying the above criteria were tagged as "*PASS*". A Fisher's exact test was performed on the read counts from

variant and reference alleles and based on the allele frequency difference the variants were classified as the following if both the tumour and germline samples had coverage: i) Germline- variants detected in both tumour and matching germline sample; ii) StrongSomatic- variants detected in tumour sample only; iii) StrongLOH- variants detected in germline sample only, opposite of StrongSomatic. For regions where only one sample had coverage, variants were classified as: i) SampleSpecific- detected in tumour sample, but no coverage in germline sample and ii) Deletion- detected in germline sample, but no coverage in tumour sample. Single nucleotide variants (SNVs) that were exclusively detected in the tumour and not in the germline DNA (tagged as "StrongSomatic") and had passed all the default *VarDict* filters (tagged as "PASS") were considered somatic SNVs (SSNVs) and selected for further analyses. The following post-filtering was performed on the SSNVs where a minimum read depth of 30X and minimum variant allele frequency of 0.2 cut-offs were applied.

2.12 Mutation signature analysis

Mutational signatures were generated using the SSNVs identified in the tumour samples using the R package *deconstructSigs* (Rosenthal *et al.*, 2016). The somatic signature profiles of tumour samples were generated using the predefined mutational signatures COSMIC (version3) (http://cancer.sanger.ac.uk/cosmic/signatures). The weights of each mutational signature contributing to the total mutational catalogue of the tumour samples were calculated by applying multiple linear regression model. The weight for each signature in *deconstructSigs* was calculated through an iterative approach which was then normalised between 0 and 1. Mutational signatures with a weight 0.06 or higher were considered significant as described in *deconstructSigs* (Rosenthal *et al.*, 2016).

Chapter 3 Genome-wide DNA methylation profile of Invasive Lobular Breast Cancer

3.1 Introduction

ILBC has been increasingly recognised as a distinct breast cancer subtype with unique morphological features, clinical presentation and response to therapy (section 1.5). However, the underlying biological explanation for many of these differences in clinical behaviour is still unknown and their impact on current clinical decision-making is limited.

Examination of expression profiles of breast cancers has advanced work in molecular subtyping. Recent studies have highlighted the differentiating molecular features between ILBC and non-ILBC by investigating the molecular features of the tumours. The somatic genomic alterations unique to ILBC tumourigenesis has been reviewed in detail in section 1.5.4. ILBC has also been shown to display a distinct gene expression profile. Comparing the transcriptomic profiles of ILBC (n=21), IDBC (n=38), two samples with lymph node metastasis and three normal breast samples, Zhao et al., (2004) reported that ILBC showed differential expression at genes involved in cell adhesion and mobility, lipid and fatty acid transport and metabolism, immune response and electron transport (Zhao *et al.*, 2004). A similar difference in transcriptomic profiles of ILBC and IDBC was also reported by (Bertucci *et al.*, 2008). Oliveira et al., (2016) compared the proteomic profiles of ILBC and IDBC and reported a differential protein levels that involved structural proteins, metabolic enzymes, molecular chaperones/heat-shock proteins, binding and transport proteins (Oliveira *et al.*, 2016). Identifying unique

molecular alterations related to ILBC tumourigenesis and progression, presents opportunities for precision, prevention and precision medicine.

In contrast to the attention put to expression profiling approaches to breast cancer subtyping, DNA methylation alterations specific to ILBC are not well characterised. Studies focusing on ILBC-specific DNA methylation alterations are mainly based on candidate gene approaches and have reported gene-specific methylation alterations (section 1.8). To the best of our knowledge, there is no study that has investigated the genome-wide DNA methylation differences between ILBC and non-ILBC tumours. There has also been a growing interest in the promising applicability of tumour DNA methylation in breast cancer prognostication (section 1.7.1). However, there is no study that has specifically investigated the DNA methylation profiles of ILBC tumours and its association with disease prognosis.

Given this gap in knowledge, we sought to study the tumour DNA methylation landscape of ILBC in detail. We first used a candidate gene approach to assess the methylation patterns of ILBC at the genes that have previously been reported to have an altered methylation pattern in ILBC that is presented in Part I: Candidate gene approach. Secondly, we investigated the genome-wide DNA methylation profiles of ILBC tumours and tested the following hypotheses: i) ILBC have a distinct genome-wide DNA methylation profile as compared to non-ILBC, that is presented in the section Part II: Genome-wide DNA methylation pattern of ILBC and ii) Genome-wide variation in DNA methylation patterns within ILBC reflect different tumour biologies and can be used as a prognostic biomarker, as presented in the section Part III: Association of variably methylated tumour DNA regions with overall survival for ILBC.

3.2 Method overview

3.2.1 Study participants and data

Analyses in this chapter included 492 invasive breast cancer samples. Details of the samples and data being used in this chapter and the corresponding sections in the thesis where detailed information is available are summarised in Table 3.1.

Table 3.1: Study participants and data.

Result	Part I: Candidate gene approach	Part II: Genome-wide DNA methylation pattern of ILBC	Part III: Association of variably methylated tumour DNA regions with overall survival for ILBC
Sample	ILBC (n = 151) Non-ILBC (n = 341) Adjacent normal breast samples (n = 13)	ILBC (n = 151) Non-ILBC (n = 341)	ILBC (n = 130) ILBC, TCGA dataset (n = 168)
Total	Tumour (n = 492) Normal (n = 13)	Tumour (n = 492)	Tumour MCCS (n = 130) Tumour TCGA (n = 168)
	A detailed clinical and pathological description of the study participants is presented in Table 2.1 (section 2.1) of the thesis. Details of TCGA data download is presented in section 2.8		
Study	MCCS kConFab ABCFR	MCCS kConFab ABCFR	MCCS TCGA
	A description of the study design is presented in section 2.1 of the thesis.		
Sample type	FFPE Tumour enriched DNA and adjacent normal DNA (Details of sample preparation is presented in section 2.3.1. Data from normal adjacent DNA has been previously described in (Wong *et al.*, 2016).		
Data information	Genome-wide DNA methylation using Illumina HM450K array (Details of the methylation assay is presented in section 2.5). Gene expression data (Normalised counts, RNA sequencing-Illumina Hi-Seq) was downloaded from TCGA (Details presented in section 2.8).		

ILBC: Invasive lobular breast cancer. Non-ILBC: Non-lobular invasive breast cancer. MCCS: Melbourne Collaborative cohort Study. kConFab: The Kathleen Cuningham Foundation Consortium for Research into Familial Breast Cancer. ABCFR: Australian Breast Cancer Family Registry. TCGA: The Cancer Genome Atlas. FFPE: Formalin-fixed paraffin embedded. HM450K: Illumina HumanMethylation 450K array.

3.2.2 Statistical analyses specific to part III.

3.2.2.1 Endpoints

Incidences of cancer cases and deaths in the MCCS participants are regularly updated by linkage to the Victorian and National cancer and death registries, which are considered to be virtually complete. The latest linkage was completed on 31 March 2017 and death data were considered to be complete up to 31 December 2016. Overall survival was defined as the time (in years) from breast cancer diagnosis to death (from any cause) or end of follow-up.

3.2.2.2 Survival analysis

Survival analyses were undertaken for the ten most variably methylated regions identified across the MCCS ILBC samples. Follow-up started at the date of diagnosis and ended at the date of death or end of follow-up, whichever came first. Cox proportional hazards regression models were used to calculate HRs and 95% CI for the association between DNA methylation levels (M values) and risk of death. Three models were fitted: i) univariable, with DNA methylation as a crude predictor; and multivariable ii) with additional adjustment for age at diagnosis and iii) with adjustment for age at diagnosis and tumour stage. For each VMR, the methylation level was defined as the average methylation value across all CpG sites covering the VMR. The same analysis was carried out using the 168 ILBC samples from TCGA. Survival analyses were undertaken using the R package *Survival* (Therneau, 2014). HRs from the two individual studies were then pooled using fixed-effects meta-analysis with inverse variance weights.

3.2.2.3 Association with gene expression

To test if DNA methylation correlated with gene expression at the ten strongest VMRs (identified in the MCCS) we assessed the correlation between average methylation levels (average M-values for all CpGs covering a VMR) and gene expression levels using Pearson's correlation; we used matching gene expression and DNA methylation data available in the TCGA dataset for nine of the ten strongest VMRs. The correlations with gene expression were also assessed for individual CpG sites of each VMR.

3.3 Results

3.3.1 Methylation data pre-processing and normalisation

Genome-wide DNA methylation was measured in 502 invasive breast cancer cases. DNA methylation data was pre-processed and normalised using the Functional normalisation (FNORM) method (section 2.6). The data quality was first evaluated by assessing the detection *P*-value. After normalisation, CpG probes with a mean detection *P*-value of more than 0.05 in one or more samples were considered unreliable and were removed from further analysis. After filtering out the poor-quality probes, we were left with a total of 449,005 CpG probes. Ten samples with mean detection *P*-value of more than 0.01 were considered poor quality and were removed at this stage and we were left with a total of 492 samples that included 151 ILBC and 341 non-ILBC cases (Figure 3.1).

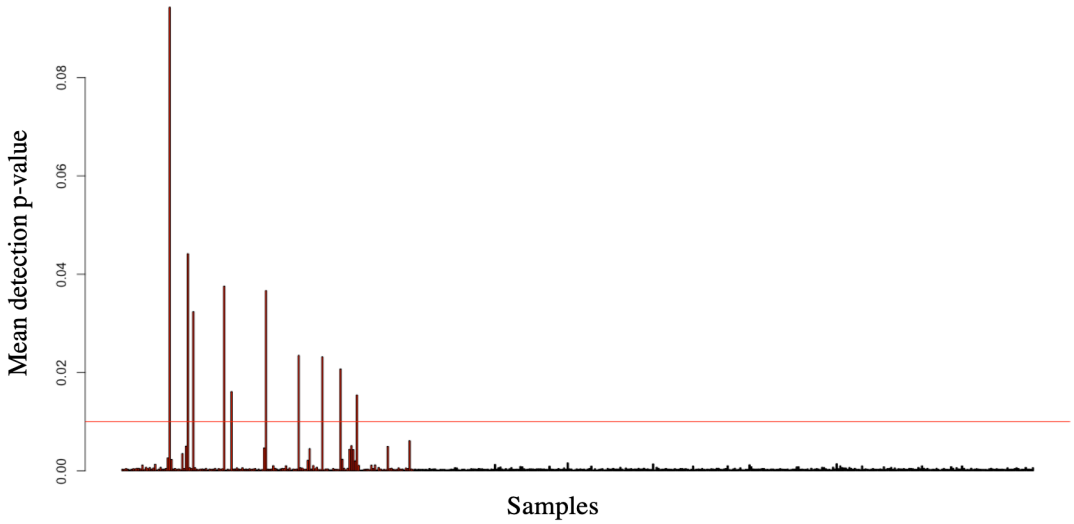


Figure 3.1: Mean detection P-value.

Plot showing mean detection *P*-value distribution for all the breast cancer samples. The samples are shown on the x-axis and the mean detection *P*-value is shown on the y-axis. Red line represents the mean detection *P*-value cut-off of 0.01.

Figure 3.2 shows the density distribution of beta-value before and after data normalisation for all the samples (n = 492).

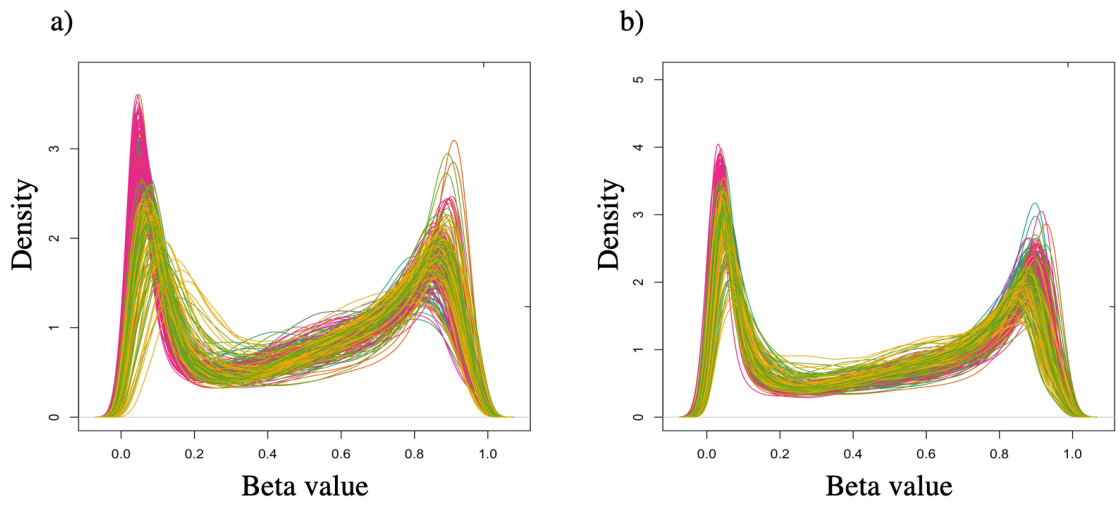


Figure 3.2: Beta-value density plots a) before data normalisation b) after data normalisation.

Density plots showing the density distribution of methylation levels (beta-value) on the x-axis and density on the y-axis. The individual samples are represented by the line in the plots.

3.3.2 Tumour purity

The tumour purity ranged from 37% to 88% across the ILBC samples with 88% of the samples showing a tumour purity of > 50%. For non-ILBC samples, the tumour purity ranged from 16% to 90% with 84% of the samples showing a tumour purity of > 50%.

Part I: Candidate gene approach

We investigated the methylation pattern of ILBC tumours at six genes; *CDH1*, *APC*, *RASSF1*, *ADAM33*, *TWIST1* and *DAPK1,* which have been previously reported to have aberrant methylation patterns in ILBC and in breast cancer overall. We also investigated the methylation pattern of ILBC tumours at the breast cancer predisposition genes; *BRCA1* and *BRCA2.* Non-ILBC samples (n=341) and the matching adjacent normal breast cancer samples (n=13) were treated as control groups in this analysis.

The methylation pattern was studied across different genomic regions in relation to the gene, which are: TSS200 - the region from transcription start site (TSS) to -200 nucleotides upstream of TSS; TSS1500 - the region from -200 to -1500 nucleotides upstream of TSS; 5 prime UTR (5'UTR) - the region within 5 prime untranslated region between the TSS and the ATG start site; 1st Exon; gene body - the region between the ATG and the stop codon and 3 prime UTR (3' UTR) - region between the stop codon and poly A signal.

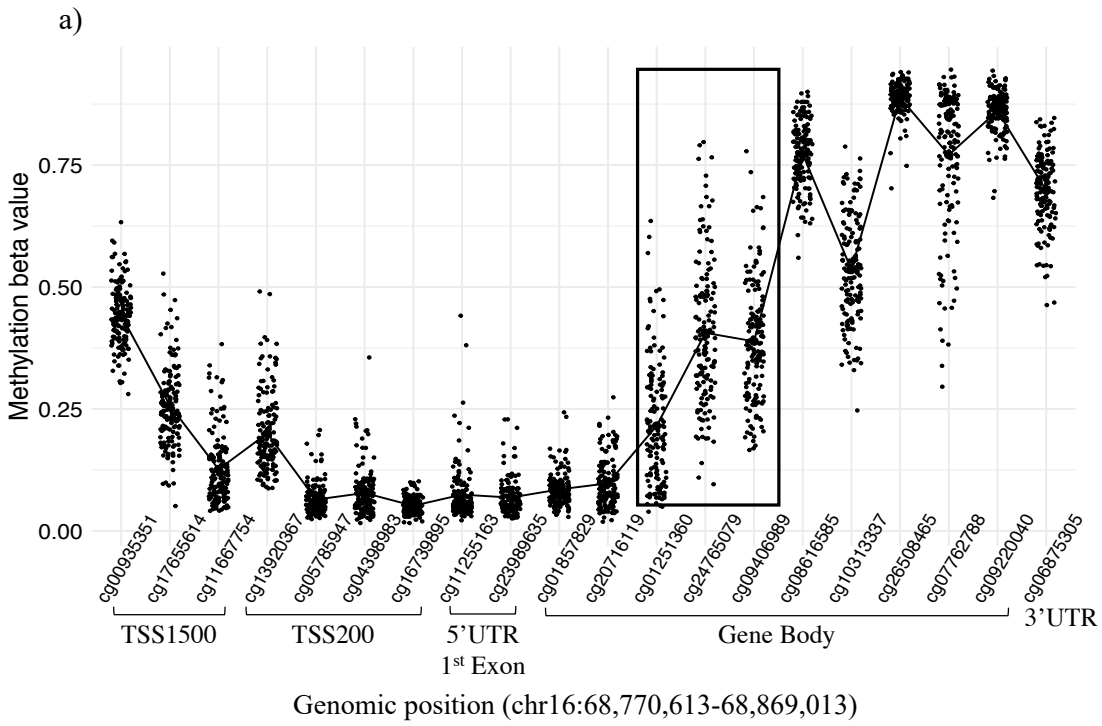
3.3.3 *CDH1*

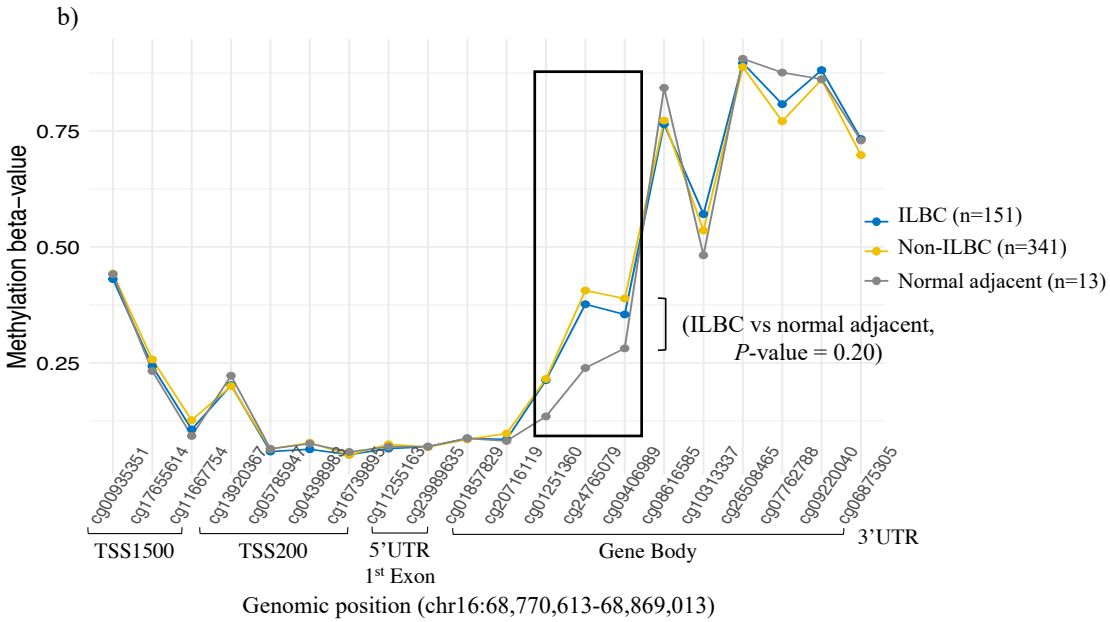
Dysfunction of *CDH1* is considered a hallmark of lobular histological subtype of breast cancer and a complete loss of the e-cadherin protein expression has been reported in ~90% of ILBC tumours (Reed *et al.*, 2015). Studies have reported *CDH1* promoter methylation as one of the mechanisms for the loss of protein expression in ILBC tumours (Droufakou *et al.*, 2001; Sarrió *et al.*, 2003).

We investigated the methylation pattern of ILBC tumours at 20 CpG positions across the genomic region, chr16:68,770,613-68,869,013, covering the *CDH1* gene. ILBC tumours showed a hypomethylation pattern across the *CDH1* promoter associated regions (TSS1500, TSS200, 5 prime UTR, 9 CpGs) with average methylation level ranging from 0.09 to 0.25 (mean = 0.15) across this region (Figure 3.3a). On the other hand, ILBC tumours were hypermethylated across the gene body and 3 prime UTR regions of *CDH1* with average methylation level ranging from 0.43 to 0.61 (mean = 0.52) (Figure 3.3a).

A variability in the DNA methylation levels (as indicated by the boxed region in (Figure 3.3a), was observed across a small region (3 CpG positions) located in the gene

body of *CDH1*. The methylation level across this region ranged from 0.11 to 0.71 across the ILBC samples. Across this region, the adjacent normal breast samples showed a lower methylation level (mean beta-value = 0.22) compared with the ILBC samples (mean beta-value = 0.34) however, the difference was not statistically significant (t.test, *P*-value = 0.20) (Figure 3.3b). Comparing the methylation patterns of ILBC, non-ILBCs and the adjacent normal samples across CDH1 promoter associated regions, we did not observe a significant difference in their methylation levels (mean beta-value, ILBC = 0.15 versus non-ILBC = 0.14, t.test, P-value = 0.89; mean beta-value, ILBC = 0.15 versus adjacent normal = 0.15, P-value = 0.94) (Figure 3.3b).





TSS200- region from transcription start site (TSS) to 200 nucleotides (nt) upstream of TSS.
TSS1500- region from 200 to 1500 nt upstream of TSS.
5'UTR- region within 5 prime untranslated region, between the TSS and the ATG start site
Gene body- region between the ATG and stop codon.
3' UTR- region between the stop codon and poly A signal.

Figure 3.3: DNA methylation pattern at *CDH1*.

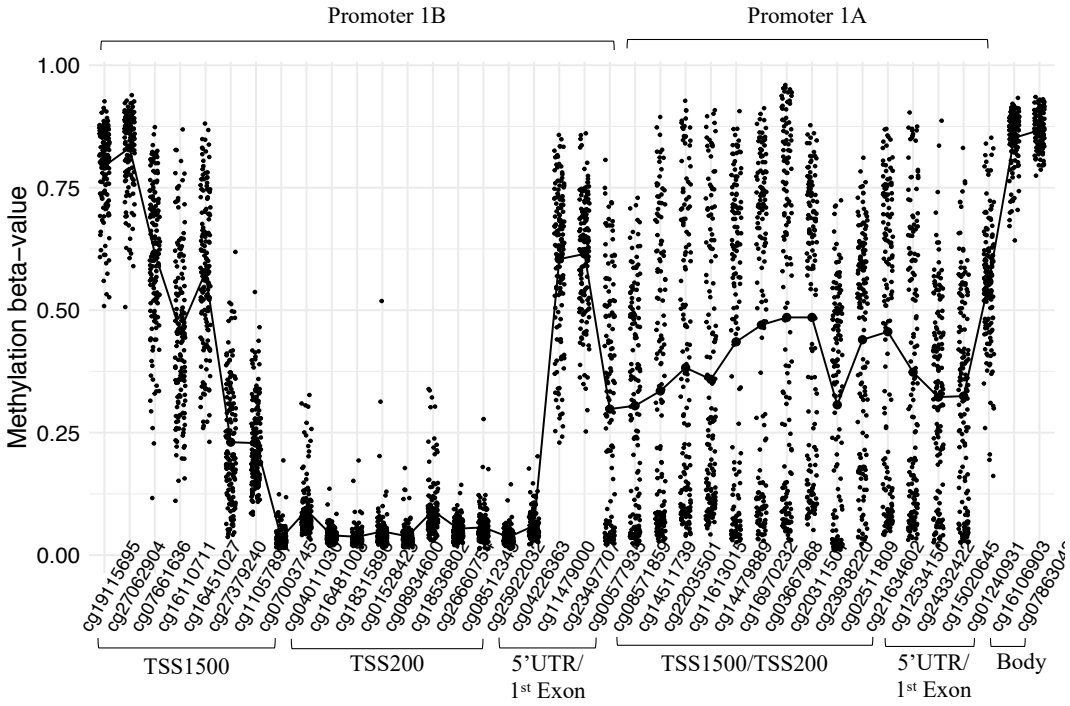
Graphics showing **a**) the methylation pattern of ILBC (n=151) across *CDH1* and **b)** the mean methylation patterns of ILBC (n=151), non-ILBC (n=341) and adjacent normal samples (n=13) across *CDH1*. The CpG positions sorted by genomic positions are shown on the x-axis and the corresponding genomic regions are marked as indicated in the legend at the bottom. The methylation level (beta-value) of the samples is shown on the y-axis. The points in plot **a** represent individual ILBC samples and the different colour lines in plot **b** represent the mean methylation levels of ILBC, non-ILBC and adjacent normal samples as indicated in the legend on the right. The region of variable methylation pattern is indicated by the boxed area. *P*-value (t.test) assessing significant difference in mean methylation levels of ILBC versus normal adjacent samples is indicated in plot **b**.

3.3.4 *APC*

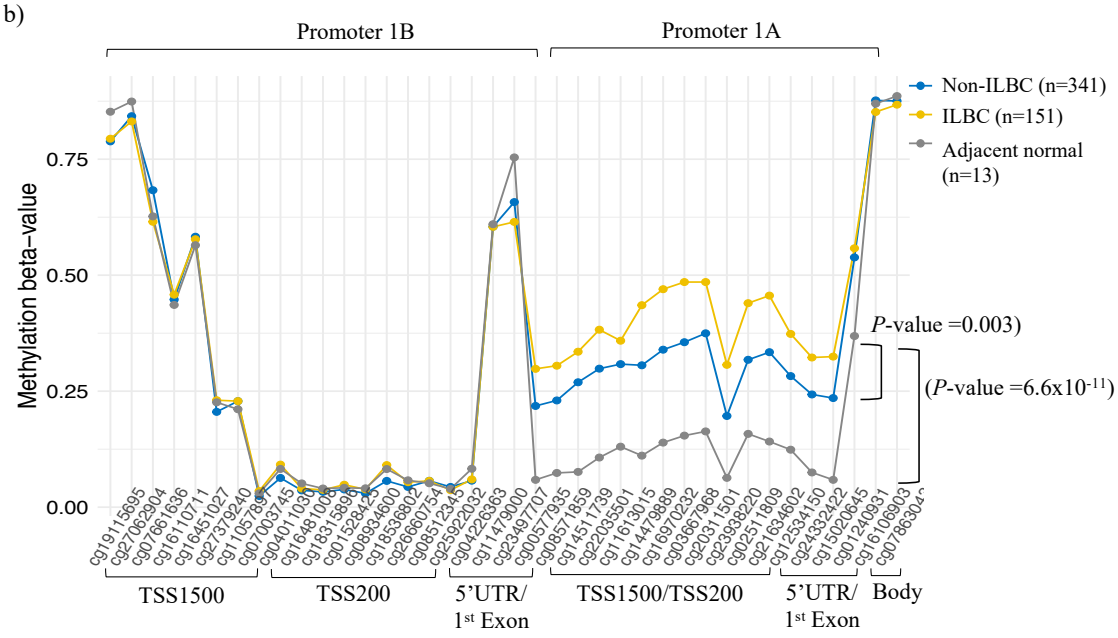
APC is a well-recognised tumour-suppressor gene and frequently reported to be hypermethylated in cancers including breast cancer (Virmani *et al.*, 2001). Two distinct promoters have been identified for *APC*, promoter 1A (Genbank accession number: U02509) and promoter 1B (accession number D13981), of which, promoter 1A is the major *APC* promoter and is reported to be most commonly active (Horii *et al.*, 1993).

We investigated the methylation pattern of ILBC tumours at 38 CpG positions across the genomic region, chr5:112,041,847-112,160,748 that spanned the two promoters of *APC* as indicated in Figure 3.4. ILBC tumours showed a substantial variability across *APC* promoter 1A (16 CpGs) with average methylation level (beta-value) ranging from 0.05 to 0.82 across this region and 66/151 (44%) ILBC tumours showing hypermethylation (Figure 3.4). On the other hand, the methylation pattern across *APC* promoter 1B (20 CpGs) was less variable with average methylation level ranging from 0.19 to 0.34 and no ILBC tumour showing hypermethylation across this region (Figure 3.4).

a)



b)



TSS200- region from transcription start site (TSS) to 200 nucleotides (nt) upstream of TSS.
TSS1500- region from 200 to 1500 nt upstream of TSS.
5'UTR- region within 5 prime untranslated region, between the TSS and the ATG start site
Gene body- region between the ATG and stop codon.
3' UTR- region between the stop codon and poly A signal.

Figure 3.4: DNA methylation pattern at *APC*.

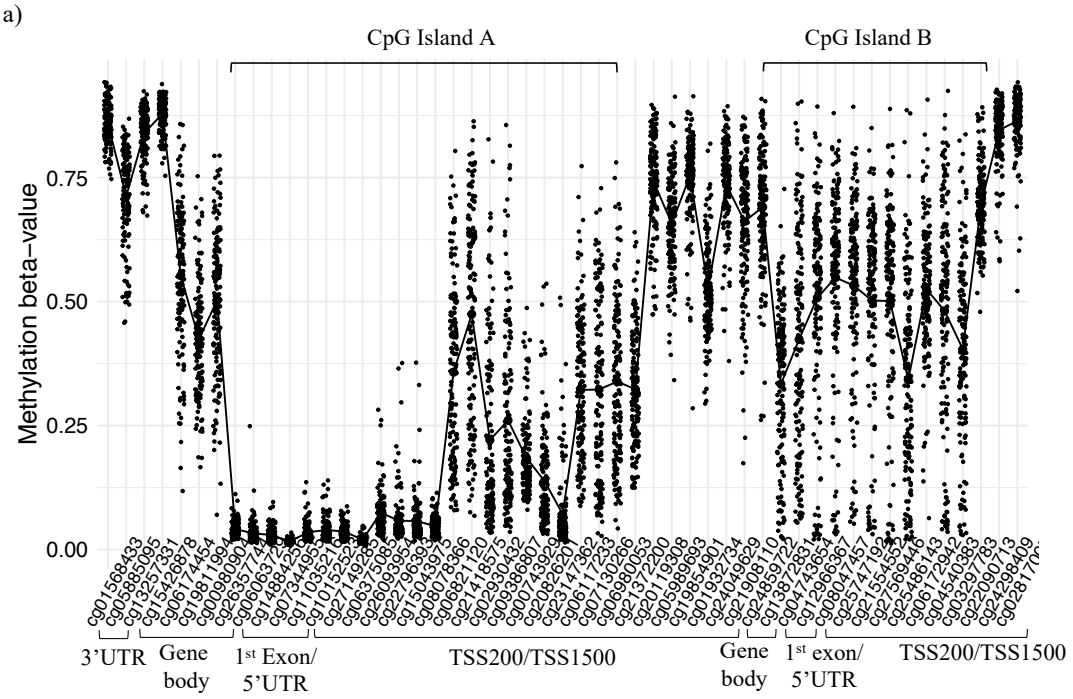
Graphics showing **a**) the methylation pattern of ILBC (n=151) across *APC* and **b)** the mean methylation patterns of ILBC (n=151), non-ILBC (n=341) and adjacent normal samples (n=13) across *APC*. The CpG positions sorted by the genomic positions are shown on the x-axis and the corresponding genomic regions are marked and indicated in the legend at the bottom. The methylation levels (beta-value) of the samples are shown on the y-axis. The points in plot **a** represent individual ILBC samples and the different colour lines in plot **b** represent the mean methylation levels of ILBC, non-ILBC and adjacent normal samples as indicated in the legend on the right. The regions associated with the two *APC* promoters are indicated. *P*-value (t.test) assessing significant difference in mean methylation level between the sample groups are indicated in plot **b**.

The methylation patterns of non-ILBC and adjacent normal samples were similar to ILBC at *APC* promoter 1B with average methylation level ranging from 0.09 to 0.38 in non-ILBC and 0.26 to 0.30 in adjacent normal breast samples across this region and no significant difference in their mean methylation levels (mean beta-value, ILBC = 0.28 *versus* non-ILBC = 0.28, t.test, *P*-value = 0.99; mean beta-value, ILBC *versus* adjacent normal breast samples = 0.29, *P*-value = 0.91) (Figure 3.4b). However, at *APC* promoter 1A, ILBC tumours showed significantly higher methylation level compared with the non-ILBC tumours (mean beta-value, ILBC = 0.40 *versus* non-ILBC = 0.30, t.test, *P*-value = 0.003; mean beta-value, ILBC *versus* adjacent normal samples = 0.13, t.test, *P*-value = $6.6x10^{-11}$) (Figure 3.4b). ILBC tumours were also more frequently methylated at *APC* promoter 1A compared with the non-ILBC tumours (66/151, 44% of ILBC tumours *versus* 97/341, 28% of non-ILBC tumours).

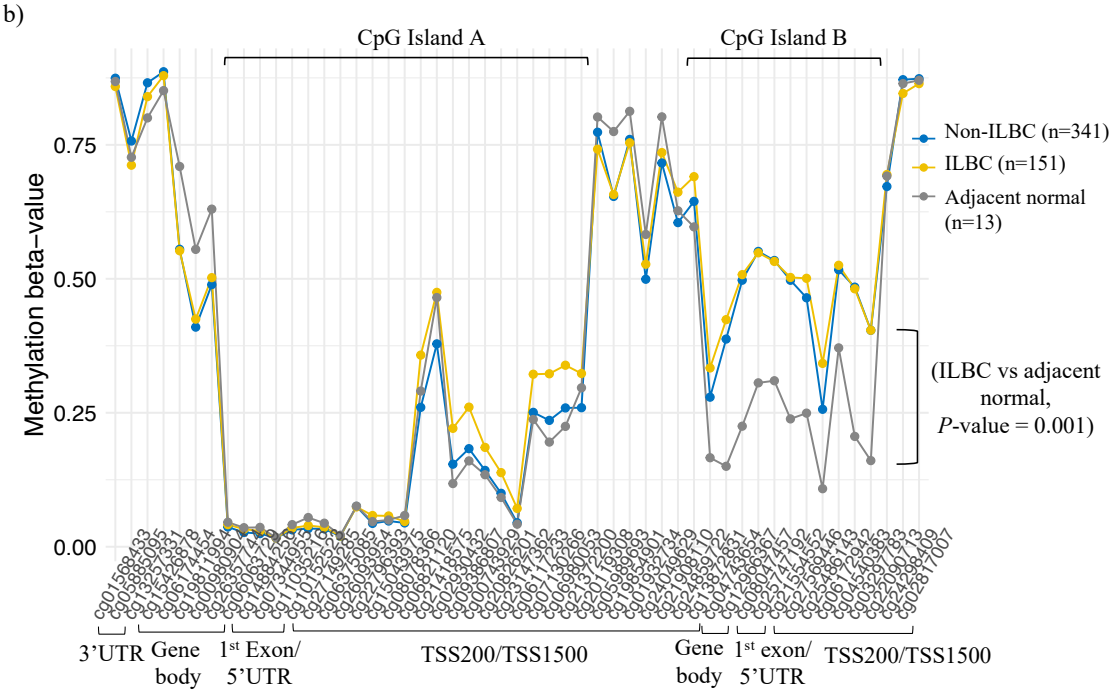
3.3.5 *RASSF1*

RASSF1 is a putative tumour suppressor gene that controls tumour growth by inhibiting the Ras pathway (Vos *et al.*, 2000). It is one of the most frequently methylated gene in cancer including breast cancer (Hesson *et al.*, 2007). Eight different transcripts *RASSF1A-RASSF1H* are known for *RASSF1*, of which transcript A and C are the most common (Agathanggelou *et al.*, 2005). Two CpG islands are associated with the promoters of *RASSF1*. The smaller CpG island spans the promoter region of *RASSF1A* (indicated as CpG island A in Figure 3.5 and the second CpG island spans the promoter region of *RASSF1B* and *RASSF1C* (indicated as CpG island B in Figure 3.5) (Agathanggelou *et al.*, 2005).

a)



b)



TSS200- region from transcription start site (TSS) to 200 nucleotides (nt) upstream of TSS.
TSS1500- region from 200 to 1500 nt upstream of TSS.
5'UTR- region within 5 prime untranslated region, between the TSS and the ATG start site
Gene body- region between the ATG and stop codon.
3' UTR- region between the stop codon and poly A signal.

Figure 3.5: DNA methylation pattern at *RASSF1*.

Graphics showing **a**) the methylation pattern of ILBC (n=151) across *RASSF1* and **b)** the mean methylation patterns of ILBC (n=151), non-ILBC (n=341) and adjacent normal samples (n=13) across *RASSF1*. The CpG positions sorted by the genomic position are shown on the x-axis and the corresponding genomic regions are marked as indicated in the legend at the bottom. The methylation levels (beta-value) of the samples are shown on the y-axis. The points in plot **a** represent individual ILBC samples and the different colour lines in plot **b** represent the mean methylation level of the ILBC, non-ILBC and adjacent normal samples as indicated in the legend on the right. The regions associated with the two *RASSF1* CpG islands are indicated. *P*-value (t.test) assessing significant difference in mean methylation level of ILBC and adjacent normal samples is indicated in plot **b**.

We investigated the methylation pattern of ILBC tumours at 51 CpG positions across the genomic region, chr3:50,367,341-50,379,334 that spanned the two CpG islands associated with *RASSF1*, indicated as CpG island A (10 CpGs) and CpG island B (21 CpGs) (Figure 3.5). The average methylation level (beta-value) ranged from 0.09 to 0.87 at CpG island A and 80/151 (53%) of ILBC tumours were hypermethylated across this region. On the other hand, the average methylation level ranged from 0.05 to 0.32 at CpG island B (21 CpGs) and no sample was found to be hypermethylated across this region (Figure 3.5a).

The methylation patterns of non-ILBC and adjacent normal samples were similar to ILBC tumours across CpG island B with no significant difference in their mean methylation levels (mean beta-value, ILBC = 0.14 *versus* non-ILBC = 0.10, t.test, *P*-value = 0.41; mean beta-value, ILBC = 0.14 *versus* adjacent normal breast = 0.11, *P*-value = 0.48) (Figure 3.5b). However, across CpG island A the tumour samples (both ILBC and non-ILBC) showed a significantly higher methylation level compared with the adjacent normal breast samples (mean beta-value, ILBC = 0.49 *versus* adjacent normal = 0.27, *P*-value = 0.001) (Figure 3.5b). No difference in the methylation levels of ILBC and non-ILBC tumours was observed across this region (mean beta-value, ILBC = 0.49 *versus* non-ILBC = 0.46, t.test, *P*-value = 0.58). However, ILBC tumours were found to be more frequently methylated across *RASSF1* CpG island A compared with non-ILBC tumours (80/151, 53% of ILBC *versus* 163/341, 48% of non-ILBC tumours).

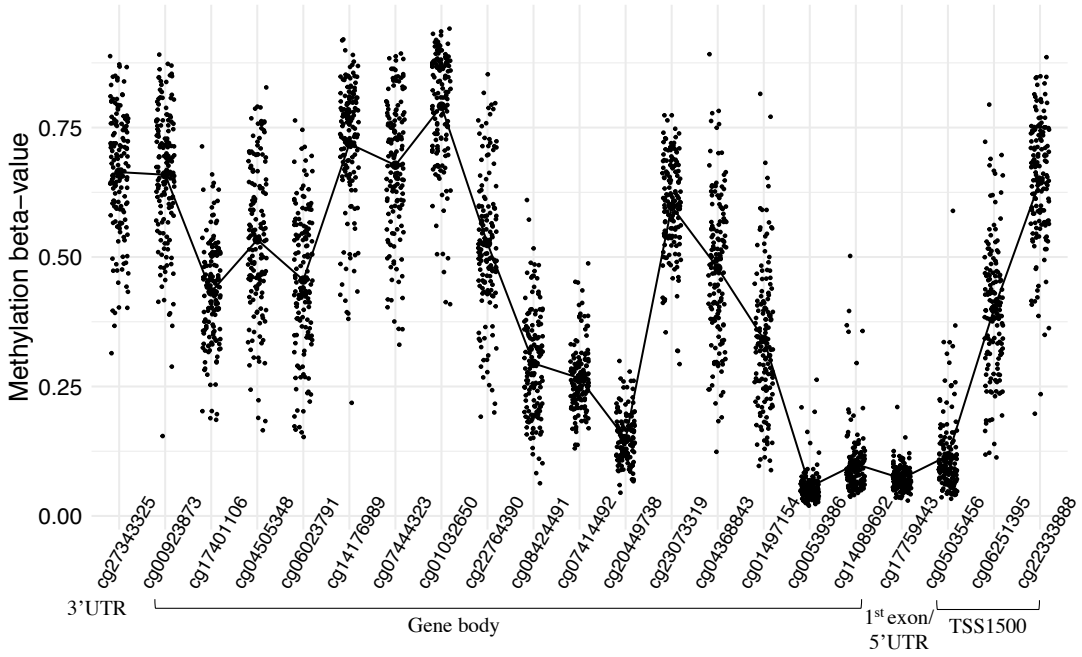
3.3.6 *ADAM33*

ADAM33 is a member of A Disintegrin and Metalloprotease (ADAM) family (Yoshinaka *et al.*, 2002) and it is involved in multiple biological functions such as proteolysis, adhesion, fusion and signalling (Stone *et al.*, 1999; Primakoff & Myles, 2000). Promoter methylation of *ADAM33* has been proposed to be a potential molecular marker for ILBC in a previous study (Seniski *et al.*, 2009).

We investigated the methylation pattern of ILBC tumours at 21 CpG positions across the genomic region, chr20: 3,649,247-3,663,421, mostly located in the gene body region (17 CpGs) and a small portion (4 CpGs), located in the promoter associated regions (TSS1500 and 1st Exon) of *ADAM33* (Figure 3.6). The average methylation level across the promoter associated regions (4 CpGs), ranged from 0.11 to 0.49 and all the samples were found to be hypomethylated (Figure 3.6a).

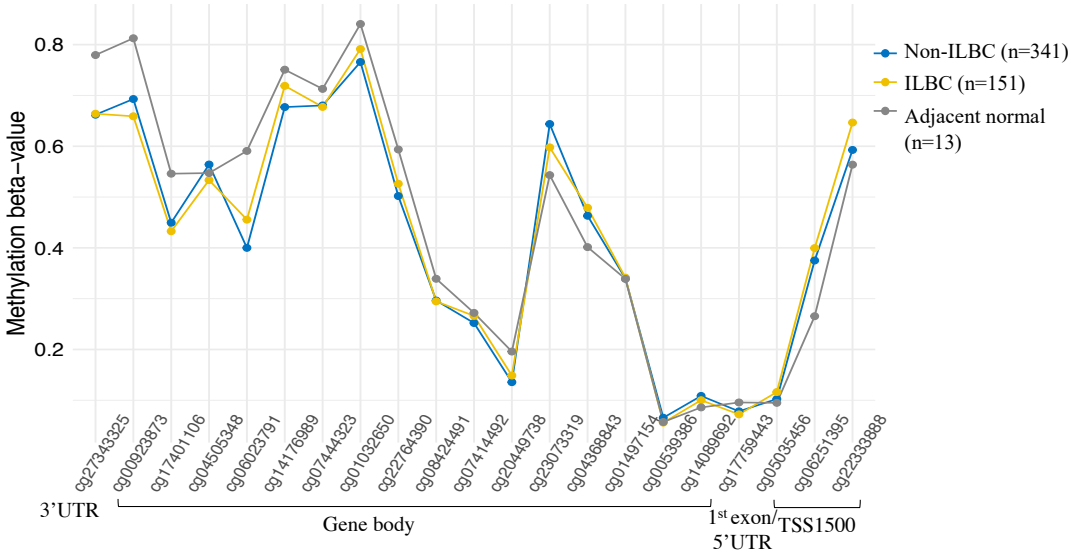
Non-ILBC and adjacent normal breast samples showed similar methylation patterns to ILBC across *ADAM33* and no significant difference in their mean methylation levels was observed across the promoter associated regions (mean beta-value, ILBC = 0.31 *versus* non-ILBC = 0.29, t.test, *P*-value = 0.91; mean beta-value, ILBC = 0.31 *versus* adjacent normal breast = 0.25, *P*-value = 0.77) (Figure 3.6b). While none of the ILBC tumours were found to be hypermethylated across *ADAM33* promoter, 5/341 (1%) non-ILBC tumours showed hypermethylation across *ADAM33* promoter region.

a)



Genomic position (chr20:3,649,247-3,663,421)

b)



Genomic position (chr20:3,649,247-3,663,421)

TSS200- region from transcription start site (TSS) to 200 nucleotides (nt) upstream of TSS.
TSS1500- region from 200 to 1500 nt upstream of TSS.
5'UTR- region within 5 prime untranslated region, between the TSS and the ATG start site
Gene body- region between the ATG and stop codon.
3' UTR- region between the stop codon and poly A signal.

Figure 3.6: DNA methylation pattern at *ADAM33*.

Graphics showing **a**) the methylation pattern of ILBC (n=151) across *ADAM33* and **b)** the mean methylation patterns of ILBC (n=151), non-ILBC (n=341) and adjacent normal samples (n=13) across *ADAM33*. The CpG positions sorted by the genomic position are shown on the x-axis and the corresponding genomic regions are marked as indicated in the legend at the bottom. The methylation level (beta-value) of the samples are shown on the y-axis. The points in plot **a** represent individual ILBC samples and the different colour lines in plot **b** represent the mean methylation level of the ILBC, non-ILBC and adjacent normal samples as indicated in the legend on the right.

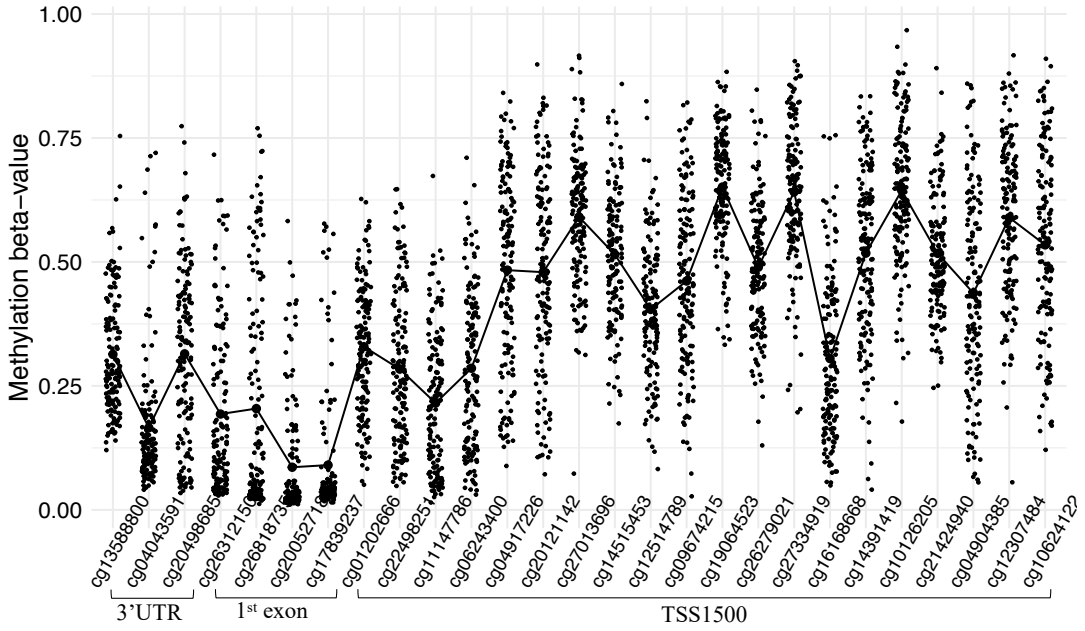
3.3.7 *TWIST1*

TWIST1 is a transcription factor, which is overexpressed in many epithelial cancers including breast cancer and is known to promote metastasis (Yang *et al.*, 2004).

The methylation patterns of ILBC tumours were investigated at 27 CpG positions across the genomic region, chr7:19,155,785-19,158,747 that mostly covered *TWIST1* promoter associated regions (TSS1500, 1st exon, 24 CpGs) and the 3 prime UTR (3 CpGs) (Figure 3.7). The average methylation level across *TWIST1* promoter regions (24 CpGs) ranged from 0.20 to 0.64 and 24/151 (16%) of ILBC tumours were found to be hypermethylated (Figure 3.7a).

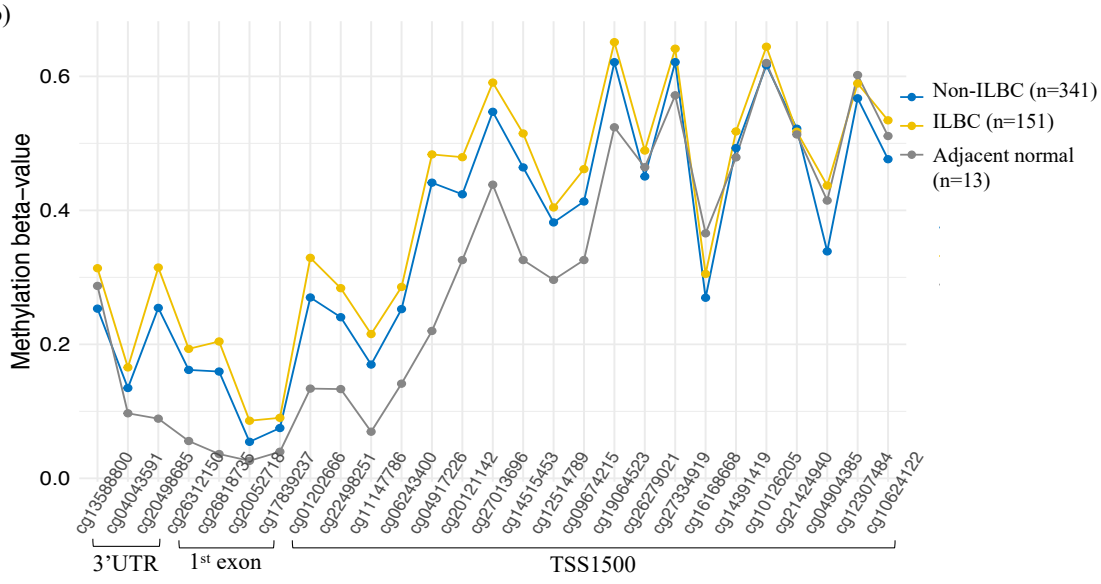
Non-ILBC and adjacent normal breast samples showed similar methylation patterns to ILBC across *TWIST1* with no significant difference in their mean methylation levels across the promoter region (mean beta-value, ILBC = 0.41 *versus* non-ILBC = 0.38, t.test, *P*-value = 0.45; mean beta-value, ILBC = 0.41 *versus* adjacent normal = 0.32, *P*-value = 0.08) (Figure 3.7b). ILBC tumours were found to be more frequently methylated at *TWIST1* promoter compared with the non-ILBC tumours (24/151, 16% of ILBCs *versus* 44/341, 13% of non-ILBCs).

a)



Genomic position (chr7:19,155,785-19,158,747)

b)



Genomic position (chr7:19,155,785-19,158,747)

TSS200- region from transcription start site (TSS) to 200 nucleotides (nt) upstream of TSS.
TSS1500- region from 200 to 1500 nt upstream of TSS.
5'UTR- region within 5 prime untranslated region, between the TSS and the ATG start site
Gene body- region between the ATG and stop codon.
3' UTR- region between the stop codon and poly A signal.

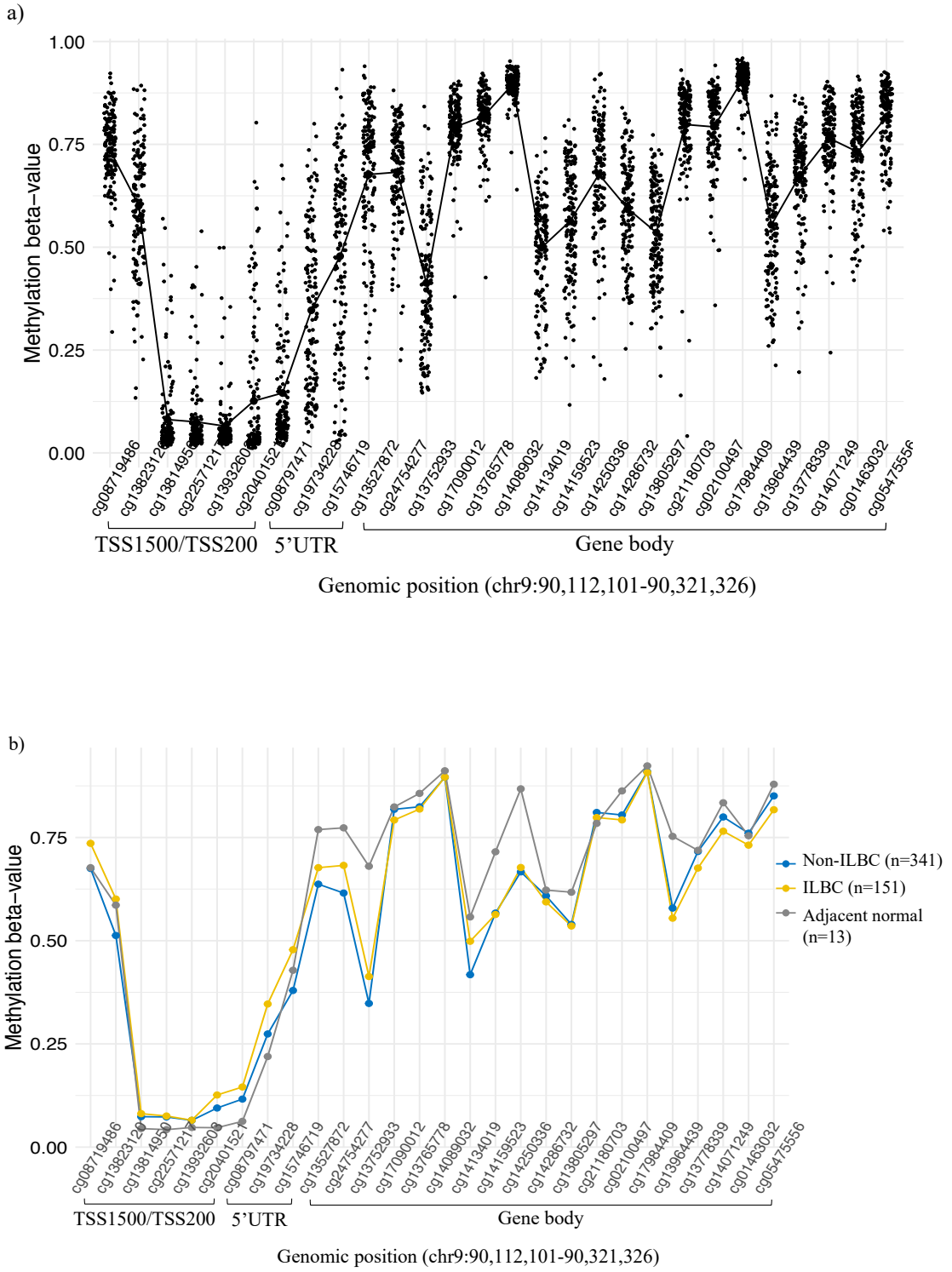
Figure 3.7: DNA methylation pattern at *TWIST1*.

Graphics show **a**) the methylation pattern of ILBC (n=151) across *TWIST1* and **b)** the mean methylation patterns of ILBC (n=151), non-ILBC (n=341) and adjacent normal samples (n=13) across *TWIST1*. The CpG positions sorted by the genomic position are shown on the x-axis and the corresponding genomic regions are marked as indicated in the legend at the bottom. The methylation level (beta-value) of the samples are shown on the y-axis. The points in plot **a** represents individual ILBC samples and the different colour lines in plot **b**) represent the mean methylation level of the ILBC, non-ILBC and adjacent normal samples as indicated in the legend on the right.

3.3.8 *DAPK1*

The *DAPK1* gene is a pro-apoptotic gene that is downregulated by promoter hypermethylation in many cancers including breast cancer (Mittag *et al.*, 2006; Jia *et al.*, 2016; Loginov *et al.*, 2017; Shawky *et al.*, 2019). Promoter methylation of *DAPK1* has been reported to be more frequent in lobular subtype compared with ductal breast cancer. A significant difference in DAPK expression level has also been reported between the two subtypes (Lehmann *et al.*, 2002).

The methylation patterns of ILBC tumours were investigated at 28 CpG positions across the genomic position, chr9:90,112,101-90,321,326 that covered the promoter associated regions (TSS1500, TSS200 and 5'UTR) and the gene body region of *DAPK1* Figure 3.8. The average methylation across *DAPK1* promoter region (9 CpGs) ranged from 0.12 to 0.62 and 4/151 (3%) of ILBC tumours were found to be hypermethylated across this region (Figure 3.8a). Non-ILBC and adjacent normal breast samples showed similar methylation patterns to ILBC across *DAPK1* promoter region with no significant difference in their mean methylation level (Figure 3.8b) (mean beta-value, ILBC = 0.30 *versus* non-ILBC = 0.25, t.test, *P*-value = 0.71; mean beta-value, ILBC = 0.30 *versus* adjacent normal = 0.24, *P*-value = 0.65). The analysis found 14/341 (4%) non-ILBC hypermethylated at *DAPK1* promoter compared with 4/151 (3%) of ILBC tumours.



TSS200- region from transcription start site (TSS) to 200 nucleotides (nt) upstream of TSS.
TSS1500- region from 200 to 1500 nt upstream of TSS.
5'UTR- region within 5 prime untranslated region, between the TSS and the ATG start site
Gene body- region between the ATG and stop codon.
3' UTR- region between the stop codon and poly A signal.

Figure 3.8: Methylation at *DAPK1*.

Graphics showing **a**) the methylation pattern of ILBC (n=151) across *DAPK1* and **b)** the mean methylation patterns of ILBC (n=151), non-ILBC (n=341) and adjacent normal samples (n=13) across *DAPK1*. The CpG positions sorted by the genomic position are shown on the x-axis and the corresponding genomic regions are marked as indicated in the legend at the bottom. The methylation level (beta-value) of the samples are shown on the y-axis. The points in plot **a** represents individual ILBC samples and the different colour lines in plot **b** represent the mean methylation level of the ILBC, non-ILBC and adjacent normal samples as indicated in the legend on the right.

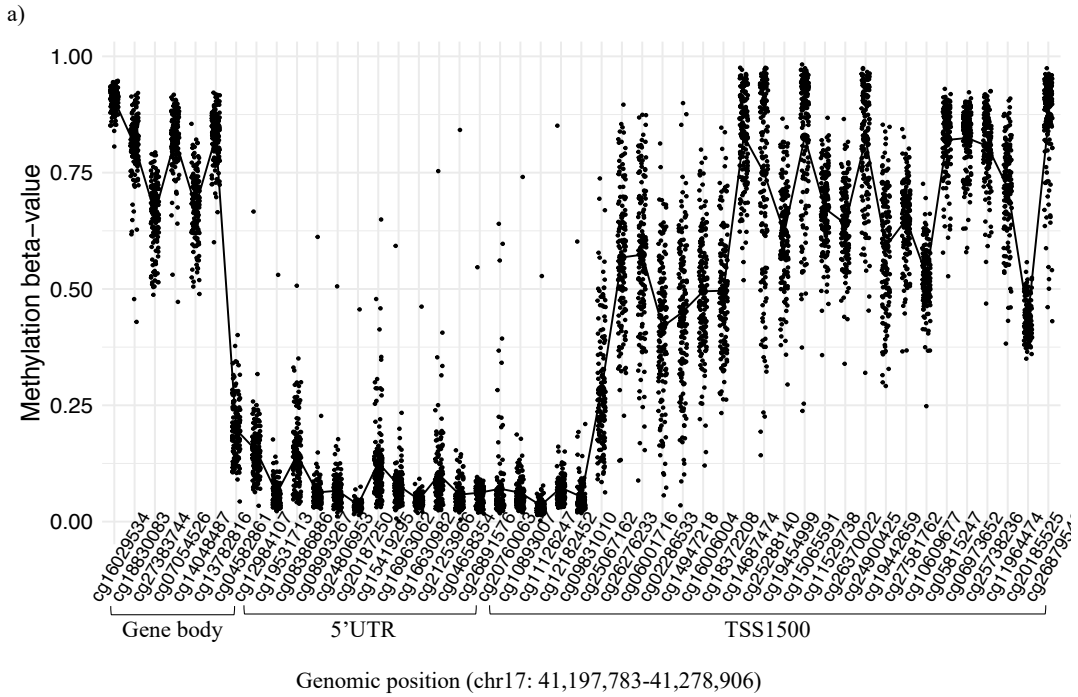
3.3.9 *BRCA1 and BRCA2*

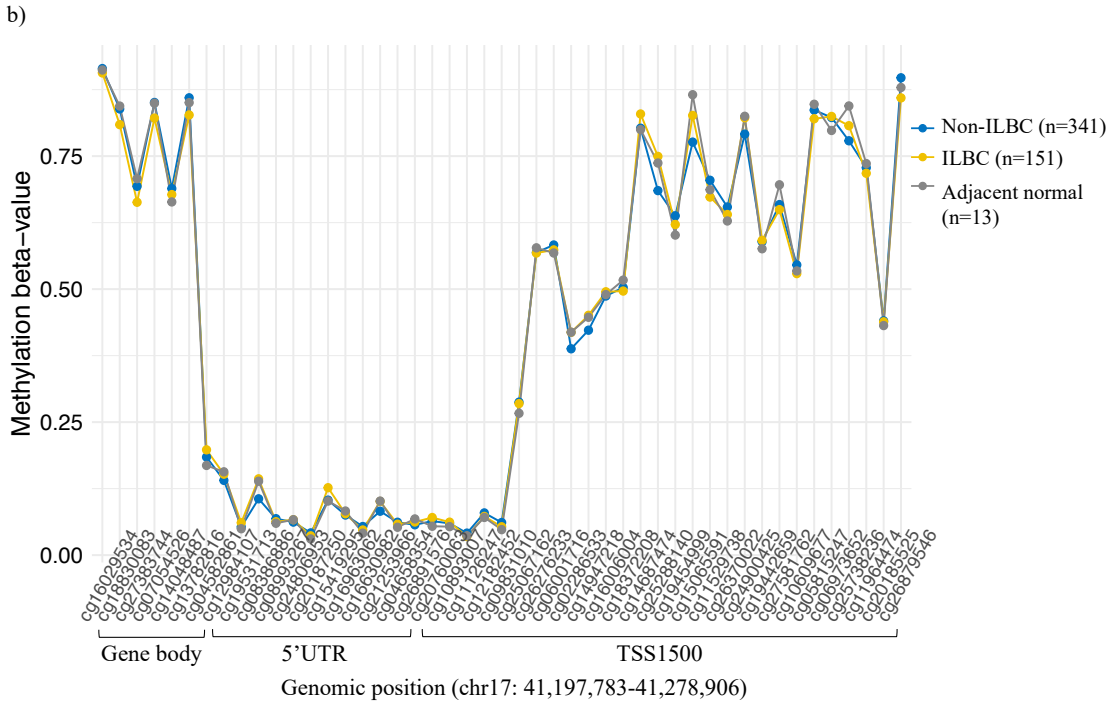
BRCA1 and *BRCA2* are breast cancer predisposition genes involved in several cellular pathways such as cell-cycle, transcriptional regulation, apoptosis and DNA repair mechanism (Roy *et al.*, 2012). The increased risk of breast cancer associated with *BRCA1* and *BRCA2* pathogenic variant has been estimated to be OR= 5.91 (95% CI:5.25-6.67) and OR= 3.31 (95% CI: 2.95-3.71), respectively (Kurian *et al.*, 2017).

Methylation patterns of ILBC tumours were investigated at 47 CpG positions across the genomic region, chr17:41,197,783-41,278,906 that covered *BRCA1* gene body and promoter associated regions (TSS1500 and 5 prime UTR) (Figure 3.9a). The promoter associated region of *BRCA1*; 5 prime UTR (11 CpGs) and TSS1500 (29 CpGs) showed different methylation patterns (Figure 3.9a). The average methylation level across 5 prime UTR was found to be consistent and hypomethylated, ranging from 0.03 to 0.60 with only one sample found to be hypermethylated across this region. The hypermethylated sample was grade III ILBC tumour with mixed lobular-ductal morphology (ICD-O code- 8522) and was negative for ER. TSS1500 on the other hand, was found to be hypermethylated across 96/151 (64%) of ILBC tumours with average methylation level ranging from 0.34 to 0.77 across this region (Figure 3.9a).

Non-ILBC tumours and adjacent normal breast samples showed similar methylation profiles to ILBC tumours across *BRCA1* promoter associated regions with no significant difference in their mean methylation levels across 5 prime UTR (mean beta-value, ILBC = 0.08 *versus* non-ILBC = 0.07, t.test, *P*-value = 0.61; mean beta-value, ILBC = 0.08 *versus* adjacent normal = 0.08, *P*-value = 0.79) and TSS1500 (mean beta-value ILBC = 0.52 *versus* non-ILBC = 0.52, *P*-value = 0.96; mean beta-value, ILBC =

0.52 *versus* adjacent normal breast = 0.52, *P*-value = 0.98) (Figure 3.9b). The analysis found 3/341 non-ILBC tumours to be hypermethylated across *BRCA1* 5 prime UTR and 197/341, 58% non-ILBC tumours to be hypermethylated across *BRCA1* TSS1500.





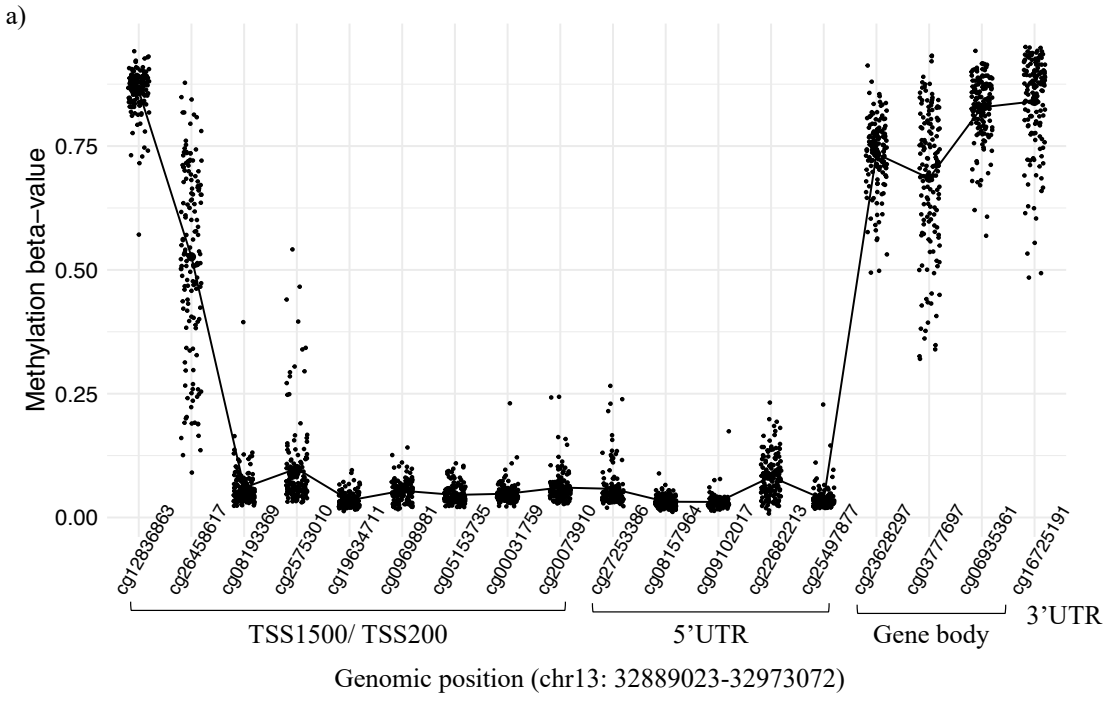
TSS200- region from transcription start site (TSS) to 200 nucleotides (nt) upstream of TSS.
TSS1500- region from 200 to 1500 nt upstream of TSS.
5'UTR- region within 5 prime untranslated region, between the TSS and the ATG start site
Gene body- region between the ATG and stop codon.
3' UTR- region between the stop codon and poly A signal.

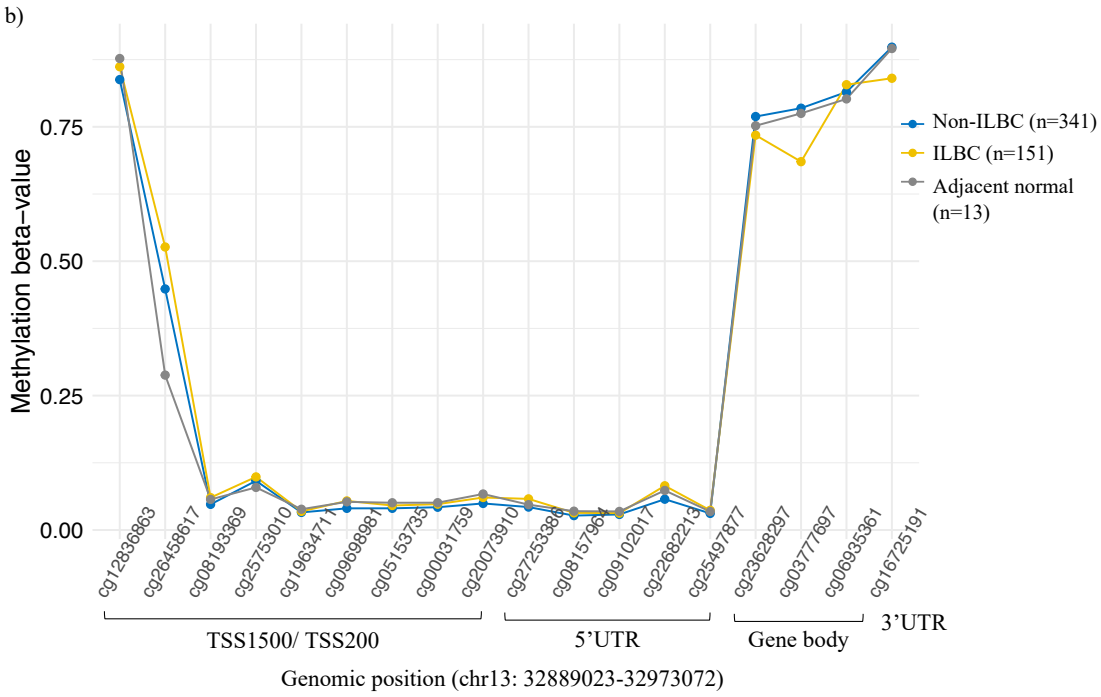
Figure 3.9: Methylation at *BRCA1*.

Graphics showing **a**) the methylation pattern of ILBC (n=151) across *BRCA1* and **b)** the mean methylation patterns of ILBC (n=151), non-ILBC (n=341) and adjacent normal samples (n=13) across *BRCA1*. The CpG positions sorted by the genomic position are shown on the x-axis and the corresponding genomic regions are marked as indicated in the legend at the bottom. The methylation level (beta-value) of the samples are shown on the y-axis. The points in plot **a** represents individual ILBC samples and the different colour lines in plot **b** represent the mean methylation levels of the ILBC, non-ILBC and adjacent normal samples as indicated in the legend on the right.

Investigation of the methylation patterns of ILBC tumours at 18 CpG positions across the genomic region, chr13:32,889,023- 32,973,072 that spanned the *BRCA2* gene body (3 CpGs), 3 prime UTR (1 CpG) and promoter associated regions (TSS1500, TSS200 and 5'UTR, 14 CpGs) (Figure 3.10).

The average methylation level across *BRCA2* promoter associated regions (14 CpGs) ranged from 0.08 to 0.22 and all the ILBC tumours were found to be hypomethylated (Figure 3.10a). Non-ILBC and adjacent normal breast samples showed a similar methylation pattern to ILBC with no significant difference in the methylation levels across *BRCA2* promoter associated regions (mean beta-value, ILBC = 0.14 *versus* non-ILBC = 0.13 t.test, *P*-value = 0.87; mean beta-value, ILBC = 0.14 *versus* adjacent normal breast = 0.13, *P*-value = 0.84) (Figure 3.10b).





TSS200- region from transcription start site (TSS) to 200 nucleotides (nt) upstream of TSS.
TSS1500- region from 200 to 1500 nt upstream of TSS.
5'UTR- region within 5 prime untranslated region, between the TSS and the ATG start site
Gene body- region between the ATG and stop codon.
3' UTR- region between the stop codon and poly A signal.

Figure 3.10: Methylation at *BRCA2*.

Graphics showing **a**) the methylation pattern of ILBC (n=151) across *BRCA2* and **b)** the mean methylation pattern of ILBC (n=151), non-ILBC (n=341) and adjacent normal samples (n=13) across *BRCA2*. The CpG positions sorted by the genomic position are shown on the x-axis and the corresponding genomic regions are marked as indicated in the legend at the bottom. The methylation level (beta-value) of the samples are shown on the y-axis. The points in **a**) represents individual ILBC samples. The different colour lines in **b**) represent the mean methylation levels of the ILBC, non-ILBC and adjacent normal samples as indicated in the legend on the right.

Part II: Genome-wide DNA methylation pattern of ILBC

To identify the genome-wide differences in DNA methylation levels between ILBC (n = 151) and non-ILBC (n = 341) tumours, a probe-wise differential methylation analysis was performed between the two groups as described in section 2.9.1.

3.3.10 Differential DNA methylation between ILBC and non-ILBC

The analysis identified 53,898 CpG positions genome-wide that were differentially methylated between ILBC and non-ILBC tumours (adjusted *P*-value < 0.01). The differentially methylated CpG positions (DMPs) were defined as the CpGs where the change in mean methylation levels (M values) of ILBC and non-ILBC differed significantly (adjusted *P*-value < 0.01). Of the DMPs, 30,869/53,898 (57%) were hypermethylated, with a positive log fold change (logFC), where logFC is the change in the average M value between the comparison groups. 23,029/53,898 (42%) of DMPs were found to be hypomethylated, with a negative logFC in ILBC in comparison with the non-ILBC tumours.

The DMPs corresponded to 13,763 genes and 8,456 intergenic regions. The most significant DMPs (adjusted *P*-value < 0.01) and the genes overlapping these DMPs were cg05968270 (chr1:65,533,502, *P*-value = 2.5×10^{-37}), cg08052428 (*RALGDS, P*-value = 3×10^{-36}), cg11658047 (*AGPAT1, P*-value = 6.5×10^{-36}), cg13286318 (*IMPAD1*, *P*-value = 1.2×10^{-35}), and cg04402633 (*CDKN1C, P*-value = 1×10^{-34}). The methylation levels of ILBC and non-ILBC samples at these DMPs are shown in Figure 3.11.

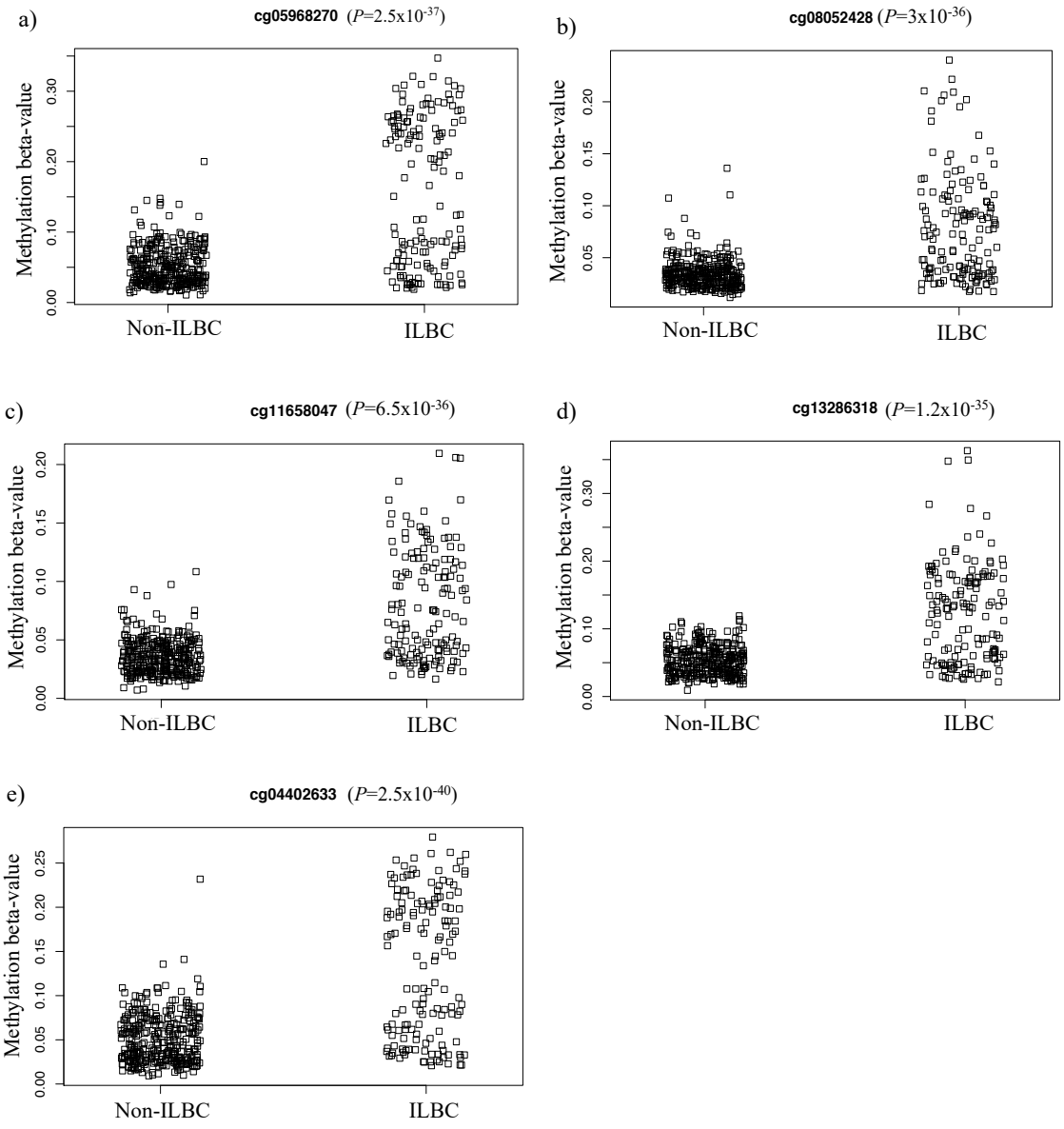


Figure 3.11: Five most significant differentially methylated CpG positions (by *P*-value) between ILBC and non-ILBC tumours.

Graphics showing the most significant differentially methylated CpG positions (DMPs); **a**) cg05968270, **b**) cg08052428, **c**) cg11658047, **d**) cg13286318 and **e**) cg04402633 between ILBC and non-ILBC tumours. Sample groups are shown on the x-axis and the methylation levels (beta-value) is shown on the y-axis. *P*-values assessing difference in methylation levels are indicated for each DMP.

To characterise the functional genomic location of differential methylation, the observed versus expected frequencies of the DMPs overlapping a functional region in the genome were evaluated. The hypermethylated DMPs were found to be enriched in CpG island (commonly associated with the gene promoter regions) by 1.64-fold. An enrichment by 1.12-fold for N-shore (up to 2 kb upstream from CpG island) and by 1.26-fold for S-shore region (up to 2 kb downstream from CpG island), in the hypermethylated DMPs was also observed (Figure 3.12a). On the other hand, the hypomethylated DMPs were found to be enriched in N-Shelf (2-4 kb upstream from CpG island) and S-shelf (2-4 kb downstream from CpG island) regions by 1.78-fold and 1.81-fold, respectively. A small enrichment (1.18-fold) in the open sea (region more than 4 kb away from the CpG island) was also observed in the hypomethylated DMPs (Figure 3.12a).

In relation to the gene, hypermethylated DMPs were enriched in the promoter associated regions; TSS200 by 1.87-fold, TSS1500 by 1.47, 5 prime UTR by 1.35-fold and in 1st exon by 1.65-fold. On the other hand, the hypomethylated DMPs were enriched in the 3 prime UTR and gene body region by 2.38 and 1.34-fold, respectively (Figure 3.12b).

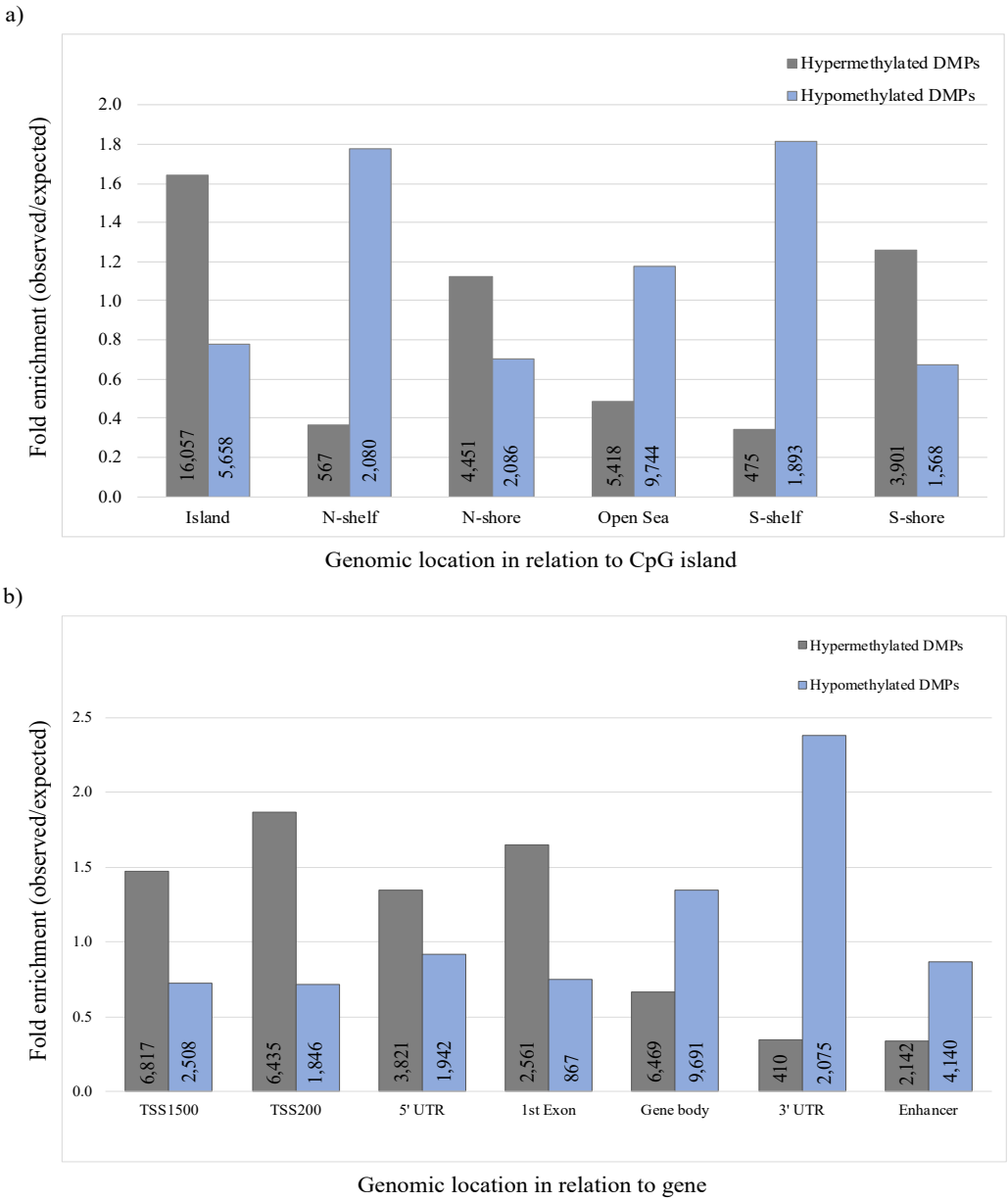


Figure 3.12: Genomic distribution of differentially methylated positions.

Bar plots showing the distribution of 53,898 differentially methylated positions (DMPs); 30,869 hypermethylated and 23,029 hypomethylated between ILBC and non-ILBC tumours **a**) relative to CpG islands, shores (0-2 kb from island), shelves (2-4 kb from island) and open sea and **b**) in relation to the gene, TSS1500 (the region from -200 to -1500 nt upstream of TSS), TSS200 (region from transcription start site (TSS) to -200 nucleotides upstream of TSS), 5'UTR (region within 5 prime untranslated region between the TSS and the ATG start site), 1st Exon, Gene body (region between the ATG and the stop codon), 3'UTR (region between the stop codon and poly A signal) and region associated with enhancers. Different genomic locations are shown on the x-axis and the fold change measured as a ratio between the frequency of hypermethylated or hypomethylated DMPs overlapping a genomic location over the expected frequency if such overlaps were to occur at random in the genome is shown on the y-axis.

Differentially methylated regions (DMRs) were identified using the *DMRcate* (Peters *et al.*, 2015) package in R as described in section 2.9.1. The DMPs (n = 53,898) collapsed into 9,543 DMRs (75% hypermethylated and 25% hypomethylated). Many significant DMRs overlapped with genes previously reported in breast cancer including *PTEN* (15CpGs, $P = 4.5x10^{-148}$, rank = 16), *MYC* (11CpGs, $P = 9.4x10^{-129}$, rank = 24), *TP53* (9CpGs, $P = 5.4x10^{-93}$, rank = 104) and *APC* (6CpGs, $P = 2.7x10^{-67}$, rank = 438), whereas many DMRs were overlapping genes that are not already known to be involved in breast cancer development. The ten most significant DMRs ($P < 0.01$) between ILBC and non-ILBC are summarised in Table 3.2.

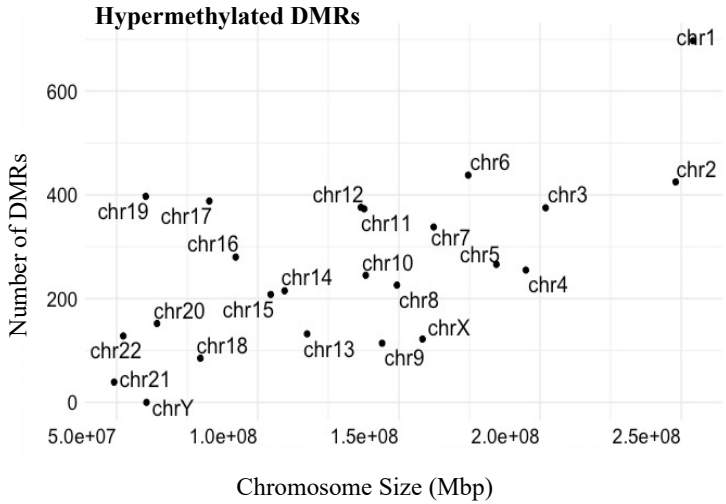
Table 3.2: Ten most significant (by *P*-value) differentially methylated regions between ILBC and non-ILBC.

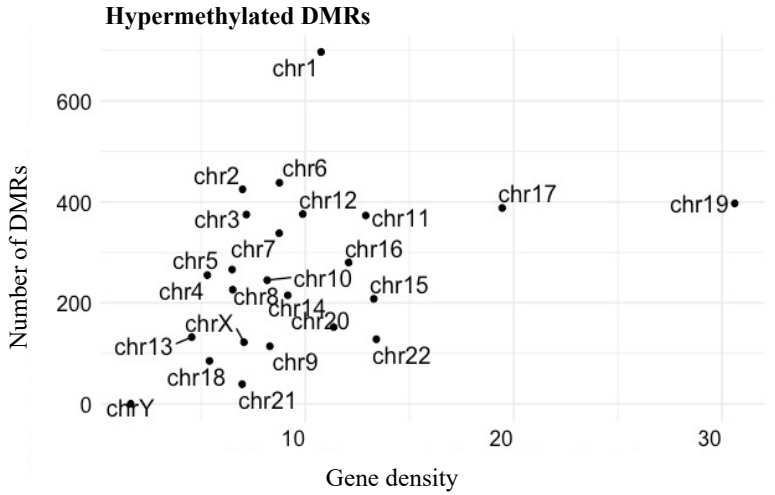
Genomic location of DMR[a] (hg19)	*P*-value*	Number of CpGs	Associated Gene[†]	Protein Function	Reported relevance to cancer progression/ etiology	Reference
chr6:32937267-32942358	0	36	*BRD2*	Transcription factor involved in cyclin gene transcription	Metabolic inflammation in breast cancer.	(Andrieu *et al.*, 2018)
chr6:31632722-31635946	$7.1x10^{-239}$	21	*Y_RNA.248, CSNK2B, GPANK1*	*CSNK2B*- Wnt signalling pathway, Regulation of *TP53* activity. *GPANK1*- Nucleic acid binding.	-	-
chr6:30710373-30712559	$1.8x10^{-238}$	21	*IER3, FLOT1*	PI3K/AKT pathway activation	-	-
chr6:30687942-30690567	$9.2x10^{-199}$	19	*TUBB*	Structural molecule activity	Hypermethylation of *TUBB* is associated with cisplatin resistant in cancers.	(Chang *et al.*, 2010)
chr6:28889996-28893092	$5.9x10^{-193}$	17	*TRIM27*	Negative regulation of interleukin-2 secretion	Promotes breast tumour growth	(Xing *et al.*, 2020)
chr6:33385325-33387188	$8.2x10^{-174}$	17	*SYNGAP1, CUTA*	*SYNGAP1*- Inhibitory regulator of the Ras-cAMP pathway; *CUTA*- Enzyme binding	-	-
chr6:30613282-30616867	$9.6x10^{-170}$	15	*C6orf136*	Uncharacterised protein	-	-
chr6:28863395-28864820	$1.7x10^{-161}$	17	*HCG14*	RNA gene	-	-
chr6:31865274-31866286	$3.1x10^{-161}$	15	*EHMT2*	Histone H3-K27 methylation, negative regulation of G0 to G1 transition.	Promotes metastasis in breast cancer.	(K Kim *et al.*, 2018)
chr5:180670343-180671419	$9.9x10^{-159}$	15	*GNB2L1*	Ribosomal protein involved in translation.	Promotes invasion and metastasis in breast cancer.	(Fan *et al.*, 2019; Buoso *et al.*, 2020)

[a] Differentially methylated region. * *P*-value computed using Stouffer's method (Stouffer *et al.*, 1949) assessing the significance of methylation difference. [†] RefSeq gene name.

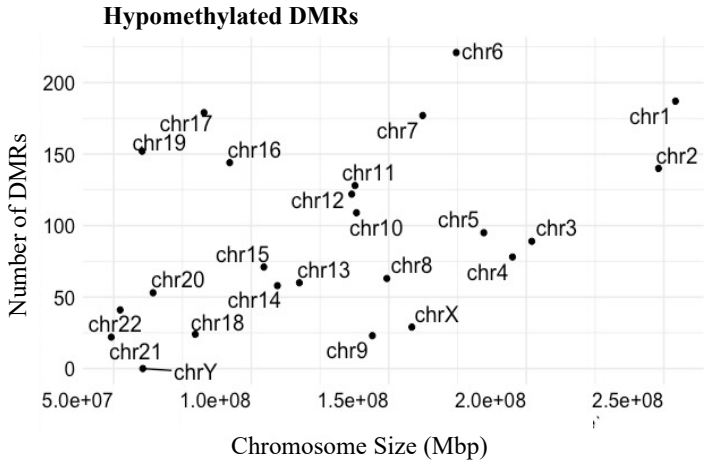
Next, the distribution of DMRs (both hypermethylated and hypomethylated) on the chromosomes were investigated to know if the DMRs were randomly distributed. A positive correlation was observed between the total number of DMRs and the chromosome length for both hypermethylated (R = 0.64, P-value = 0.0007) and hypomethylated DMRs (R = 0.43, P-value = 0.04) (Figure 3.13a). In terms of gene density (number of genes per chromosome), chromosome 1 and 6 with lower gene densities had a higher proportion of both hypermethylated (11% on chromosome 1 and 7% on chromosome 6) and hypomethylated DMRs (8% on chromosome 1 and 10% on chromosome 6) (Figure 3.13b). Overall, chromosomes 19, 17 and 16 had the highest densities of both hypermethylated and hypomethylated DMRs, whereas it was lower for chromosomes 4, 9, 18 and 21.

a)


Hypermethylated DMRs
Number of DMRs
Chromosome Size (Mbp)
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chrX
chrY
0
200
400
600
5.0e+07
1.0e+08
1.5e+08
2.0e+08
2.5e+08


Hypermethylated DMRs
Number of DMRs
Gene density
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chrX
chrY
0
200
400
600
10
20
30

b)


Hypomethylated DMRs
Number of DMRs
Chromosome Size (Mbp)
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chrX
chrY
0
50
100
150
200
5.0e+07
1.0e+08
1.5e+08
2.0e+08
2.5e+08

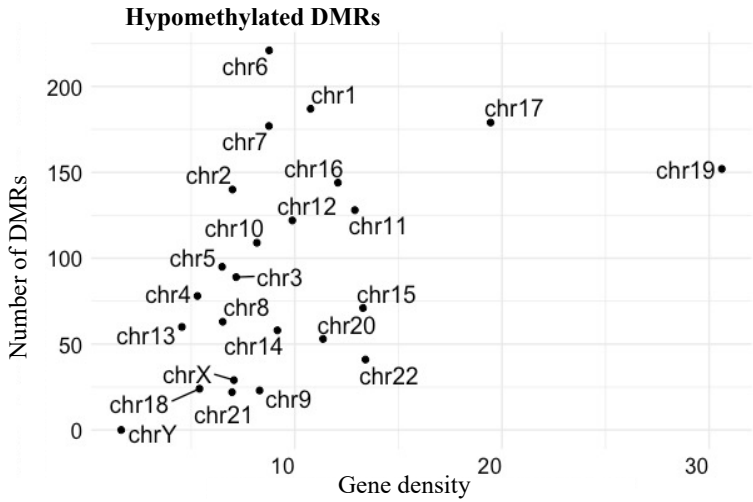

Hypomethylated DMRs
Number of DMRs
Gene density
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chrX
chrY
0
50
100
150
200
10
20
30

Figure 3.13: Relation between number of differentially methylated regions and chromosome length and gene density.

Graphics showing the relation between **a**) the number of hypermethylated differentially methylated regions (DMRs), n = 6,274 and **b**) the number of hypomethylated DMRs, n = 2,265, identified between ILBC and non-ILBC samples and chromosome length (Mbp) and gene density (number of genes per chromosome).

To further check for any hotspots of differential methylation, the DMR density using a 5 Mb sliding window was calculated. The DMRs were distributed across the genome with dense clusters at some chromosomal locations such as chromosome 19 that had hypermethylated clusters, whereas chromosome 16 and 17 showed hypomethylated DMR clusters. Chromosome 6 had intense clusters of both hypermethylated and hypomethylated DMRs (Figure 3.14)

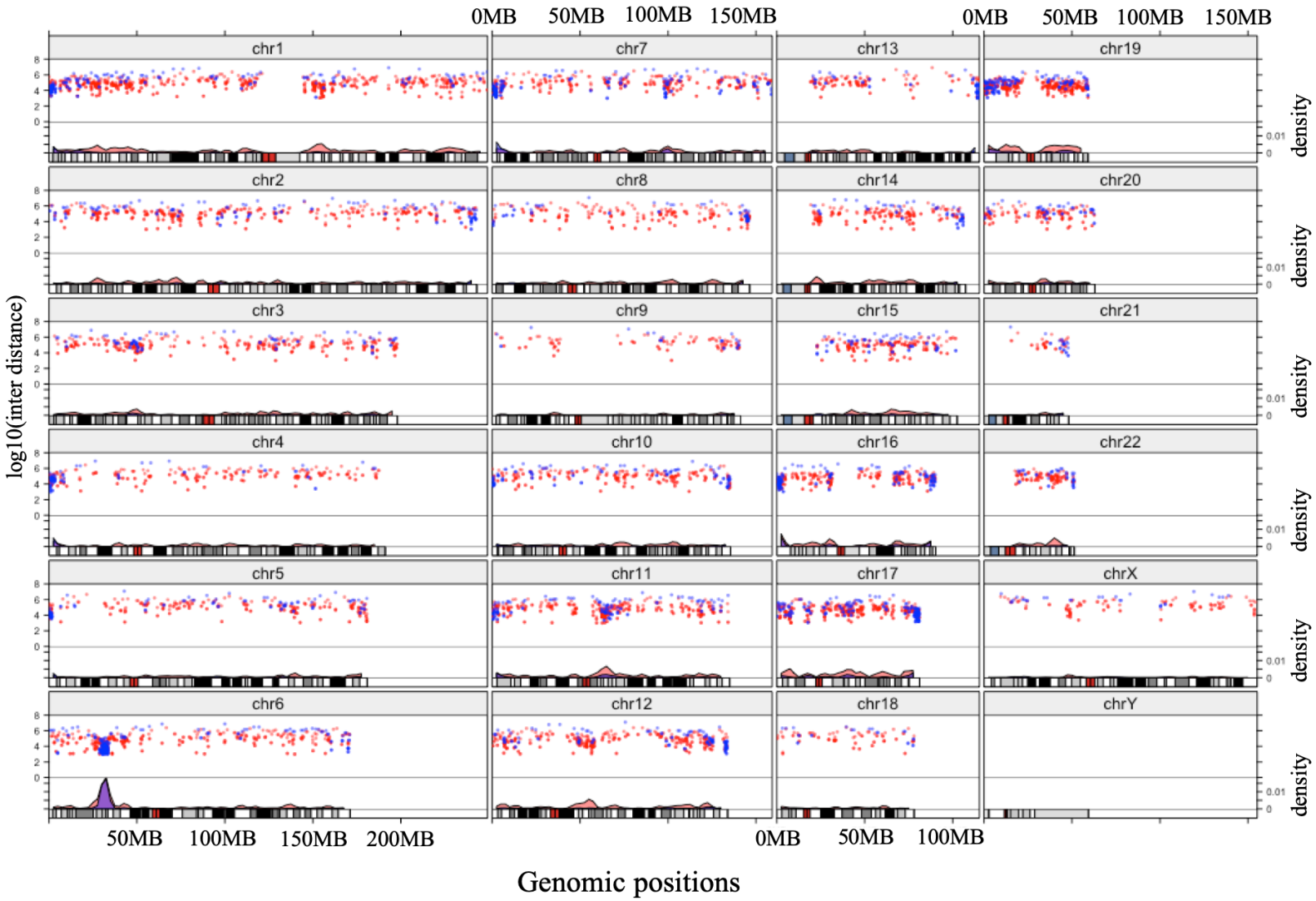

0MB
50MB
100MB
150MB
chr1
chr7
chr13
chr19
chr2
chr8
chr14
chr20
chr3
chr9
chr15
chr21
chr4
chr10
chr16
chr22
chr5
chr11
chr17
chrX
chr6
chr12
chr18
chrY
density
log10(inter distance)
50MB
100MB
150MB
200MB
Genomic positions

Figure 3.14: Differentially methylated regions between ILBC and non-ILBC.

Rainfall plot illustrating the genomic distribution of differentially methylated regions (DMRs) between ILBC and non-ILBC and also shows localized clusters of hypermethylated (n = 6,274) and hypomethylated (n = 2,265) DMRs. Four tracks are shown for each chromosome (from top to bottom): i) chromosome number; ii) rainfall plots for hypermethylated (shown in red dots) and hypomethylated (shown in blue dots) DMRs; iii) genomic density for hypermethylated (pink density plot) and hypomethylated (purple density plot) DMRs, and iv) ideograms. Each dot in the rainfall plot represents a DMR. The x-axis shows the genomic coordinate, and the y-axis shows the minimal distance (log transformed) of the DMR to its two neighbouring DMRs and the genomic density of the DMRs (defined as the fraction of a genomic window that is covered by DMRs).

3.3.11 Gene set enrichment analysis of DMRs

Gene set enrichment analysis was performed on the 1,000 most significant DMRs using *Metaspace* as described in section 2.9.3. The genes associated with the DMRs were enriched for 386 terms (FDR-adjusted $P < 0.05$) with stronger evidence for *metabolism of RNA* (R-HSA-8953854), *mRNA processing* (GO:0006397), *RNA splicing* (GO:0008380), *cell cycle* (R-HSA-1640170) and DNA repair (GO:0006281).

3.3.12 Luminal A ILBC versus Luminal A non-ILBC

DNA methylation pattern is known to be influenced by the hormone receptor and HER2 expression status of breast tumours (Widschwendter *et al.*, 2004). To minimise the molecular subtype driven heterogeneity and to better identify methylation signatures specific to ILBC an analysis limited to only luminal A samples was conducted, representing 65% of ILBC (n = 98) and 54% of non-ILBC (n = 185) cases. The samples were classified into intrinsic subtypes using immunohistochemistry (IHC) status of ER, PR and HER2. The definition reported by the St Gallen International Expert Consensus that defines luminal A as ER and/or PR positive and HER2 negative was used (Goldhirsch *et al.*, 2011). There were 36 ILBC cases for whom no IHC information was available and hence were not included in the analysis.

In the subset of only luminal A samples, 10,973 DMPs were identified out of which, 8,630 (79%) were hypermethylated and 2,343 (21%) were hypomethylated in luminal A ILBC compared with luminal A non-ILBC samples. The DMPs (n = 10,973) further clustered into 1,569 DMRs. The analysis found 498/1,569 (32%) of DMRs to be common between the two comparisons (ILBC *versus* non-ILBC and luminal A ILBC *versus* luminal A non-ILBC). Many significant DMRs in ILBC *versus* non-ILBC

comparison also ranked highly in luminal A ILBC *versus* luminal A non-ILBC comparison including, *BRD2* (*P*-value = $1.3x10^{-44}$, rank = 10), *TUBB* (*P*-value = $8.3x10^{-42}$, rank = 13), *TRIM27* (*P*-value = $1.4x10^{-41}$, rank = 15), *EHMT2* (*P*-value = $2.3x10^{-49}$, rank = 7), *GNB2L1* (*P*-value = $2.4x10^{-50}$, rank = 5) and IER3 (*P*-value = $1.0x10^{-39}$, rank= 22).

Pathway enrichment analysis of the 1,000 most significant DMRs identified between luminal A ILBC and luminal A non-ILBC showed an enrichment for 567 terms (FDR-adjusted $P < 0.05$). Of these, 283 (50%) overlapped with pathways identified in ILBC *versus* non-ILBC DMRs. Most significant pathways identified in ILBC *versus* non-ILBC DMRs were also found to be significant ($P < 0.05$) in luminal A ILBC versus luminal A non-ILBC DMRs including *metabolism of RNA* (R-HSA-8953854), *mRNA processing* (GO:0006397), *RNA splicing* (GO:0008380), *cell cycle* (R-HSA-1640170) and DNA repair (GO:0006281).

Part III: Association of variably methylated tumour DNA regions with overall survival for ILBC

Tumour DNA methylation profiling has shown potential to refine disease subtyping and improve the diagnosis and prognosis prediction of breast cancer. Here, the genome-wide variability of DNA methylation levels across a subset of ILBC tumours was investigated and the association between methylation levels at the variably methylated regions and overall survival in women with ILBC was assessed. In order to maintain the data uniformity, this analysis only included the participants from the MCCS (n = 130) and was replicated using data retrieved from TCGA for 168 ILBC cases.

3.3.13 Study participants

The median age at breast cancer diagnosis in the MCCS was 65 years with tumours being diagnosed at stage 1A/1B (65/130, 50%), 2A/2B (48/130, 37%) and 3A/3C/4 (17/130, 13%). There were 37 deaths observed during follow-up (median [interquartile range]: 13 [9-18] years). The tumours were mainly ER positive, PR positive and HER2 negative (47%). In TCGA data, the median age at diagnosis was 62 years. In both datasets, older women (aged 60 years or older at diagnosis) formed the majority of the cases (65% in the MCCS and 58% in TCGA). There was a higher proportion of young women at diagnosis (age less than 50 years: 21%) in TCGA compared with the MCCS (5%). The proportion of later stage tumours (3A/3B/3C/4) was also higher in TCGA (33%) compared with the MCCS (13%). A total of 14 deaths were recorded during the follow up (median [interquartile range]: 2 [1.5-5] years) in TCGA dataset. The clinical and pathological features of the study participants in the MCCS and TCGA and a comparison of the two studies is summarised in Table 3.3.

Table 3.3: Clinical and pathological features of the study participants from the Melbourne Collaborative Cohort Study and The Cancer Genome Atlas.

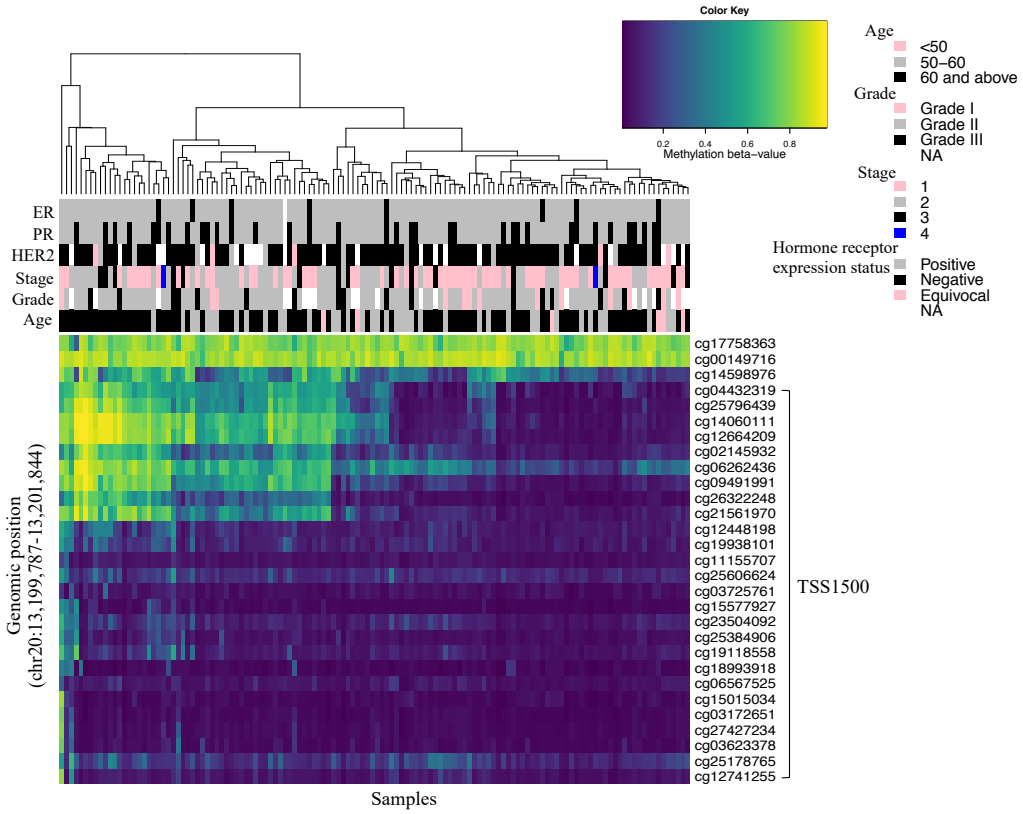
Sample characteristics	MCCS[a] (n=130)	TCGA[b] (n=168)	*P*-value*
Median age at diagnosis, years [interquartile range]	65 [25%; 58]	62 [25%; 51]	0.02
Age group, n (%)			
<50 years	6 (5)	35 (21)	0.0002
50-60 years	39 (30)	35 (21)	
60 +years	85 (65)	98 (58)	
Year of diagnosis, n (%)			
1992-1996	18 (14)	0 (0)	$4.4x10^{-30}$
1997-2001	47 (36)	4 (2)	
2002-2005	36 (28)	15 (9)	
2006 and later	29 (22)	147 (86)	
Missing	0 (0)	2 (1)	
Overall deaths, n (%)	37 (28)	14 (8)	$4.7x10^{-06}$
Median follow-up time, years	13	2	$2.2x10^{-16}$
Tumour grade, n (%)			
Grade I	13 (10)	Missing	NA
Grade II	80 (61)	Missing	
Grade III	17 (13)	Missing	
Missing	20 (15)	Missing	
Tumour stage, n (%)			
1A/1B	65 (50)	20 (12)	$1.9x10^{-12}$
2A/2B	48 (37)	92 (55)	
3A/3C/4	17 (13)	55 (33)	
Missing	0 (0)	1 (0.5)	
Tumour ER expression, n (%)			
Positive	121 (93)	157 (93)	0.32
Negative	8 (6)	6 (4)	
Missing	1 (1)	5 (3)	
Tumour PR expression, n (%)			
Positive	94 (72)	140 (83)	0.004
Negative	35 (27)	22 (13)	
Missing	1 (1)	6 (4)	
Tumour HER2 expression, n (%)			
Positive	11 (8)	21 (13)	$1.5x10^{-5}$
Negative	92 (71)	84 (50)	
Equivocal	5 (4)	35 (21)	
Missing	22 (17)	28 (17)	

[a] Melbourne Collaborative Cohort Study. [b] The Cancer Genome Atlas. ER: Estrogen receptor, PR: Progesterone receptor, HER2: Human epidermal receptor 2, **P*-values are for chi-square tests and T-tests for categorical and continuous variables, respectively.

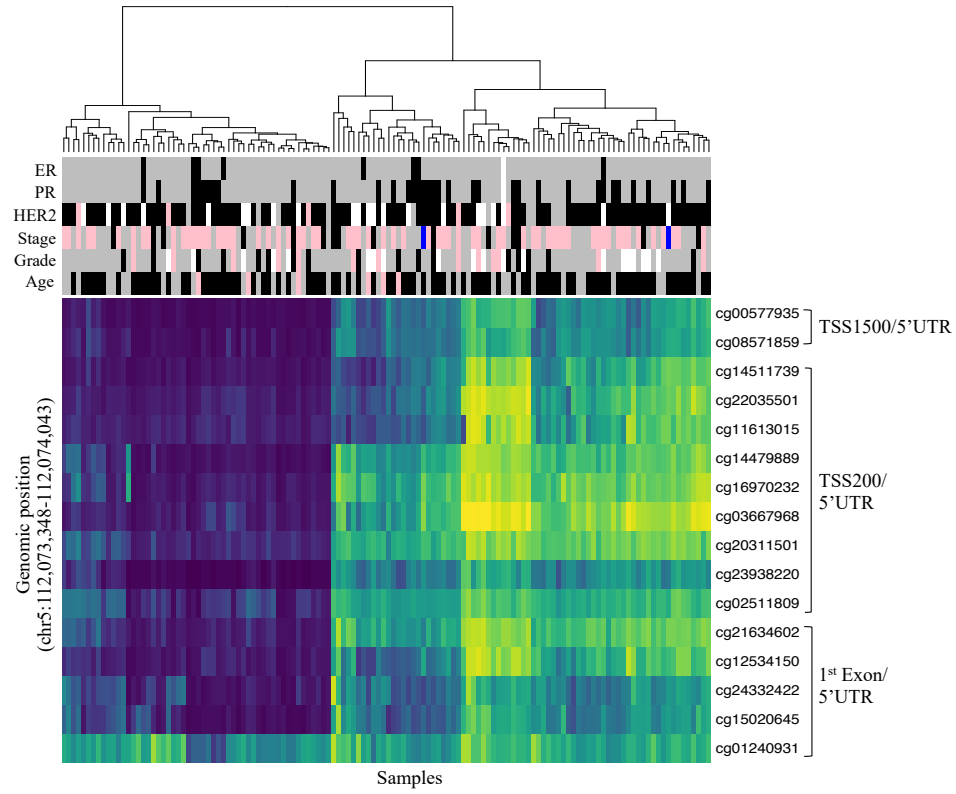
3.3.14 Variably methylated regions in ILBC

Across the genome, 2,771 regions showed substantially variable methylation ($P < 10^{-8}$) across ILBCs in the MCCS. These VMRs corresponded to 2,208 genes and 563 intergenic regions. The most significant regions ($P < 10^{-8}$) and the genes associated with these regions were chr20:13199787-13201844 (*ISM1,* 29 CpGs), chr5:112073348-112074043 (*APC*, 16 CpGs), chr17:42091713-42093050 (*TMEM101,* 16 CpGs), chr11:2290953-2293552 (*ASCL2,* 41 CpGs), chr10:134598496-134602228 (*NKX6*, 39 CpGs) and chr1:228644750-228647248 (*HIST3H2A/HIST3H2BB,* 28 CpGs). The average methylation level (beta-values) ranged between 0.09 and 0.63 at *ISM1,* 0.08 and 0.82 at *APC*, 0.15 and 0.83 at *TMEM101*, 0.15 and 0.77 at *ASCL2,* 0.07 and 0.70 at *NKX6,* and 0.05 and 0.58 at *HIST3H2A/ HIST3H2BB* (Figure 3.15).

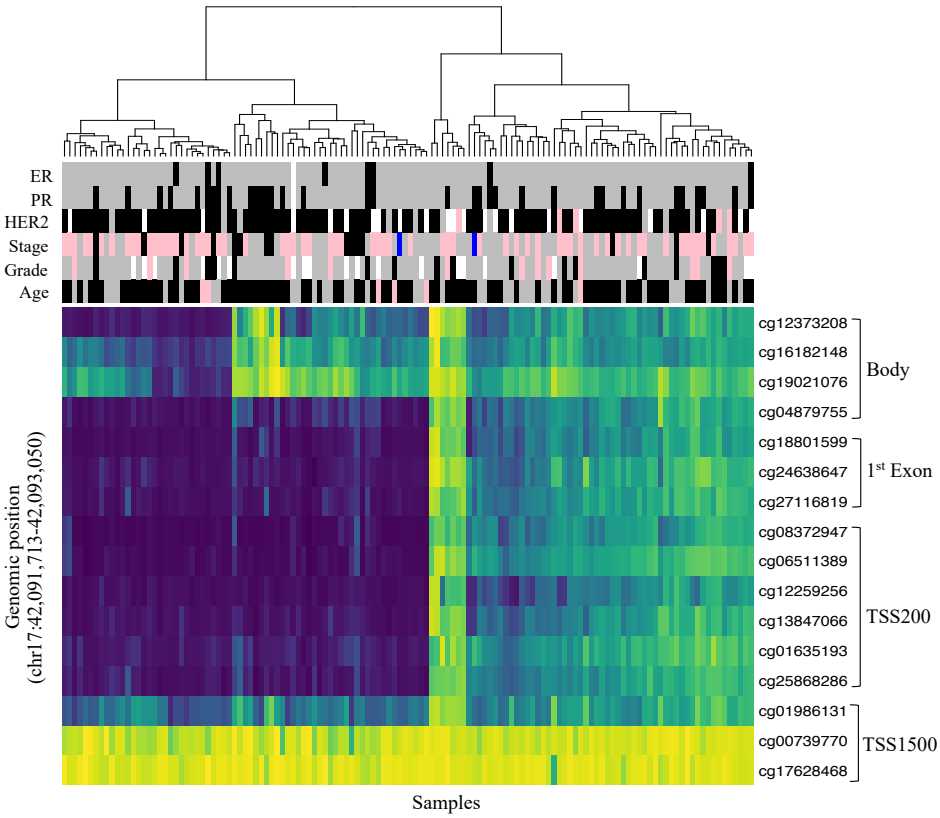
a) *ISM1*


Color Key
0.2
0.4
0.6
0.8
Methylation beta−value
Age
<50
50–60
60 and above
Grade
Grade I
Grade II
Grade III
NA
Stage
1
2
3
4
Hormone receptor expression status
Positive
Negative
Equivocal
NA
ER
PR
HER2
Stage
Grade
Age
Genomic position (chr20:13,199,787-13,201,844)
cg17758363
cg00149716
cg14598976
cg04432319
cg25796439
cg14060111
cg12664209
cg02145932
cg06262436
cg09491991
cg26322248
cg21561970
cg12448198
cg19938101
cg11155707
cg25606624
cg03725761
cg15577927
cg23504092
cg25384906
cg19118558
cg18993918
cg06567525
cg15015034
cg03172651
cg27427234
cg03623378
cg25178765
cg12741255
TSS1500
Samples

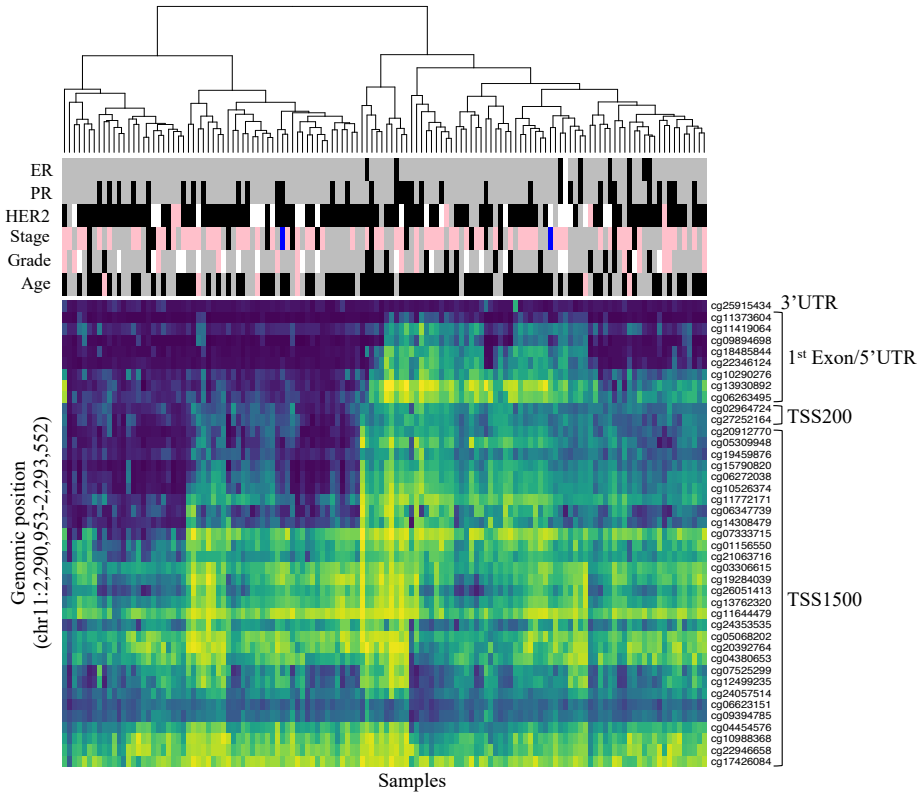
b) *APC*


ER
PR
HER2
Stage
Grade
Age
Genomic position (chr5:112,073,348-112,074,043)
cg00577935
cg08571859
TSS1500/5'UTR
cg14511739
cg22035501
cg11613015
cg14479889
cg16970232
cg03667968
cg20311501
cg23938220
cg02511809
TSS200/ 5'UTR
cg21634602
cg12534150
cg24332422
cg15020645
cg01240931
1st Exon/ 5'UTR
Samples

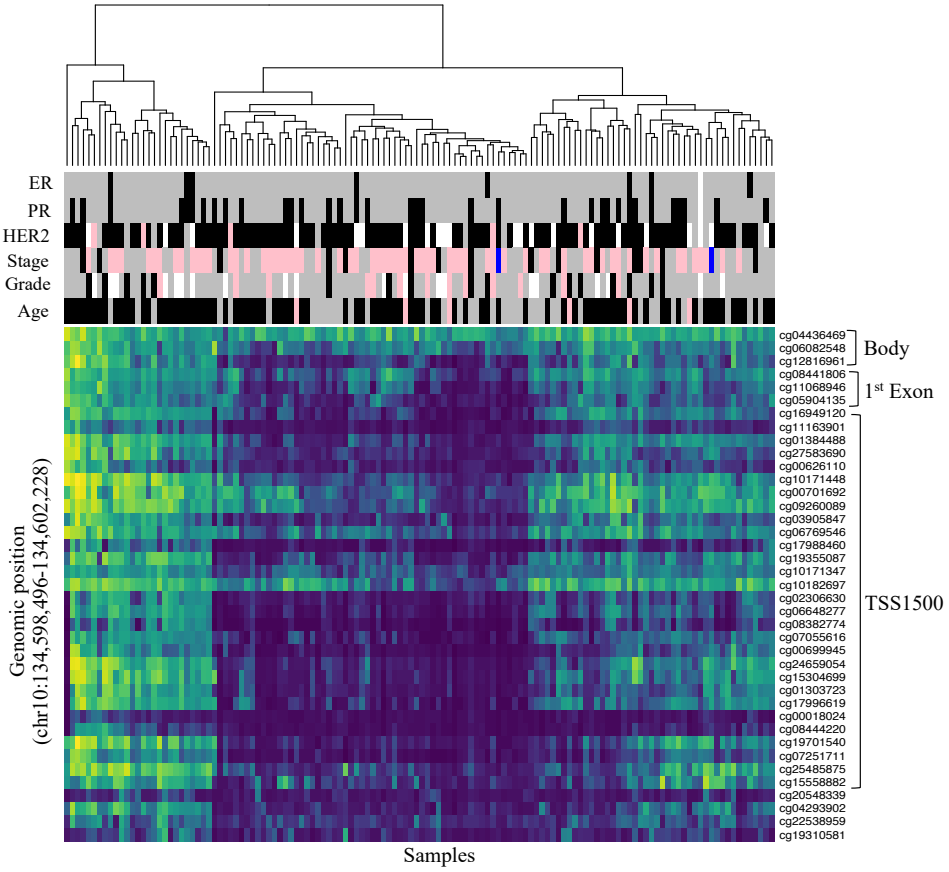
c) *TMEM101*



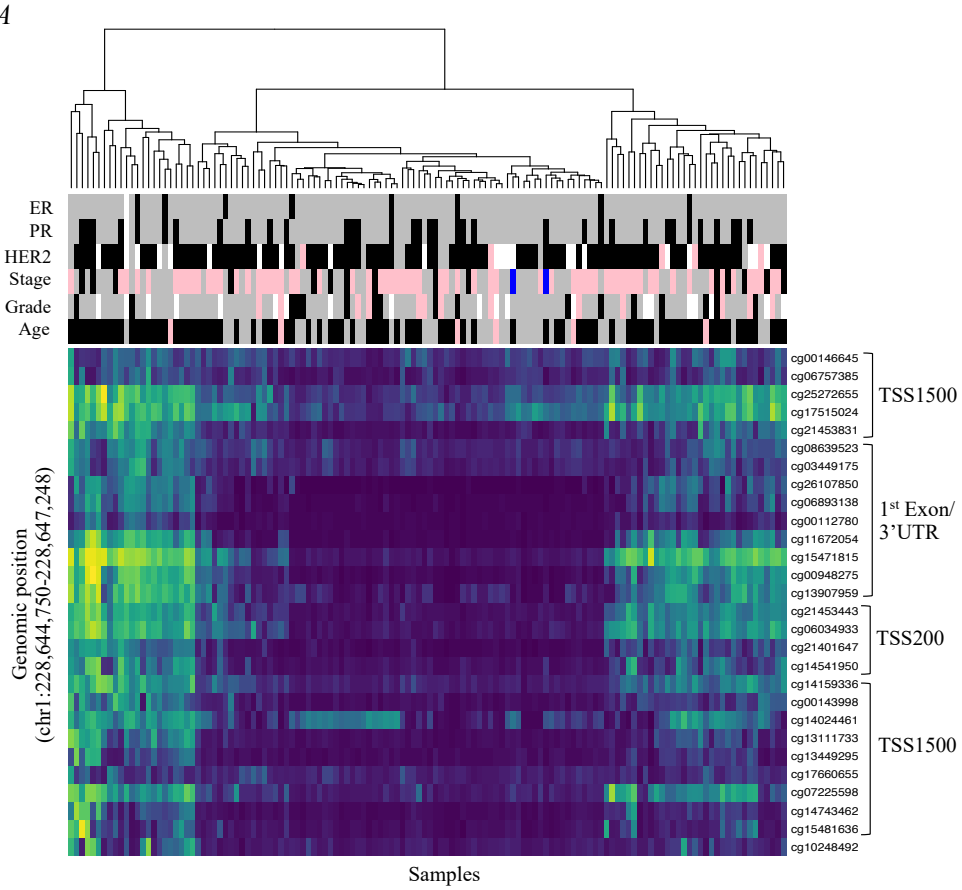
d) *ASCL2*



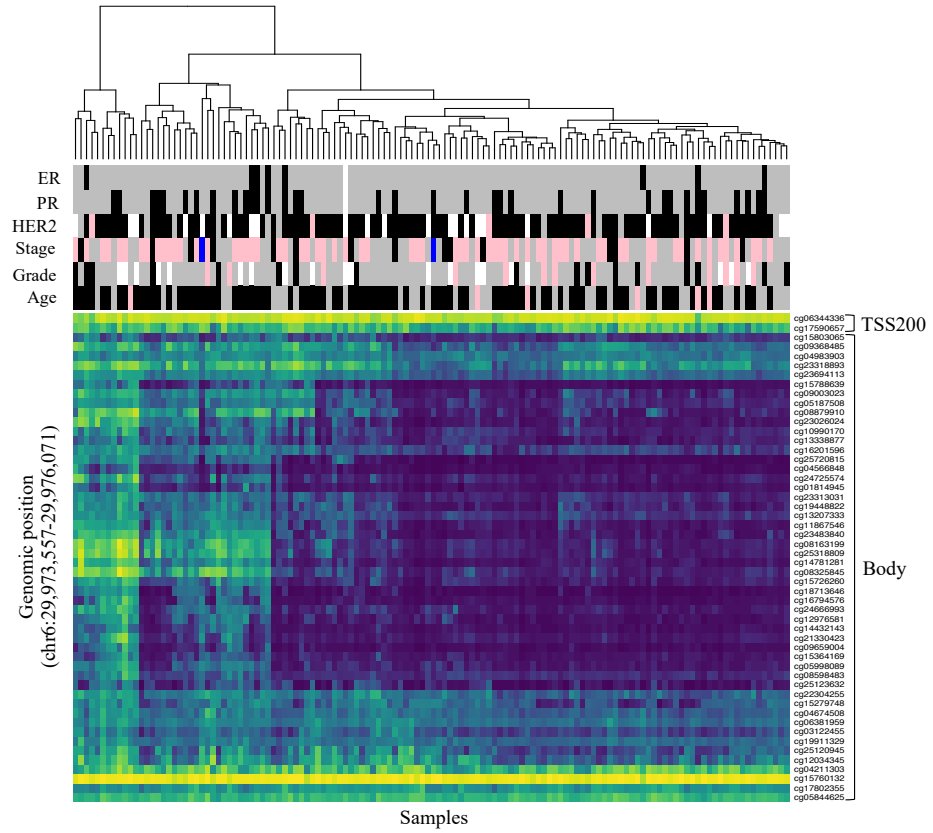
e) *NKX6*



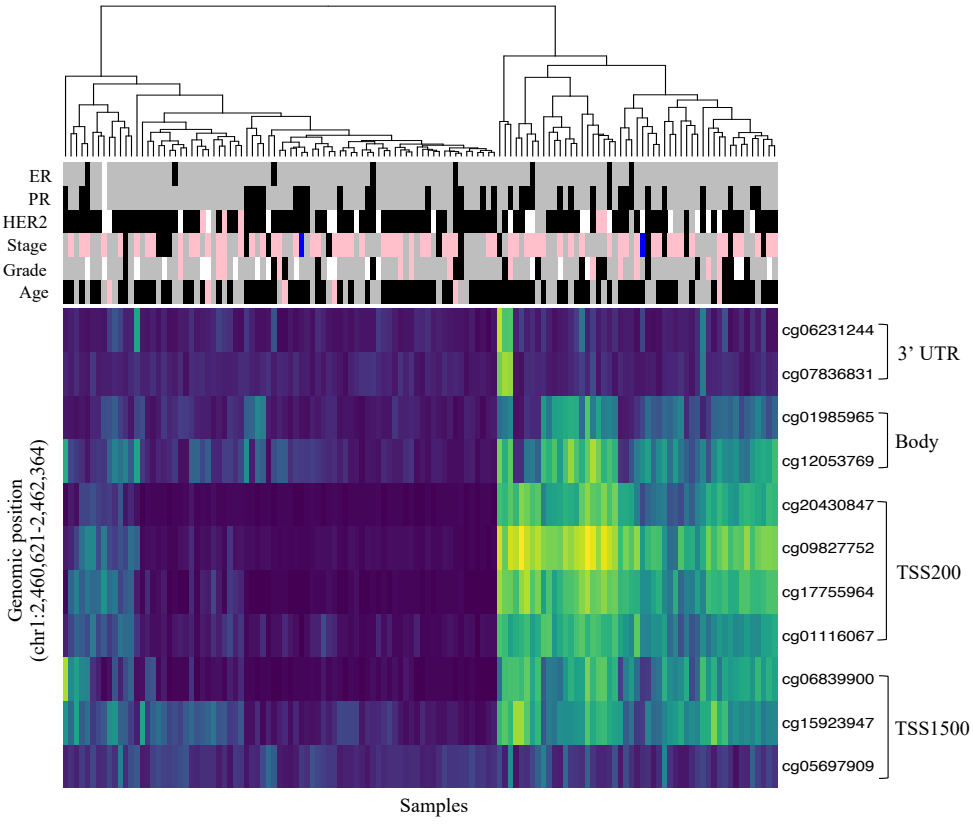
f) *HIST3H2A*



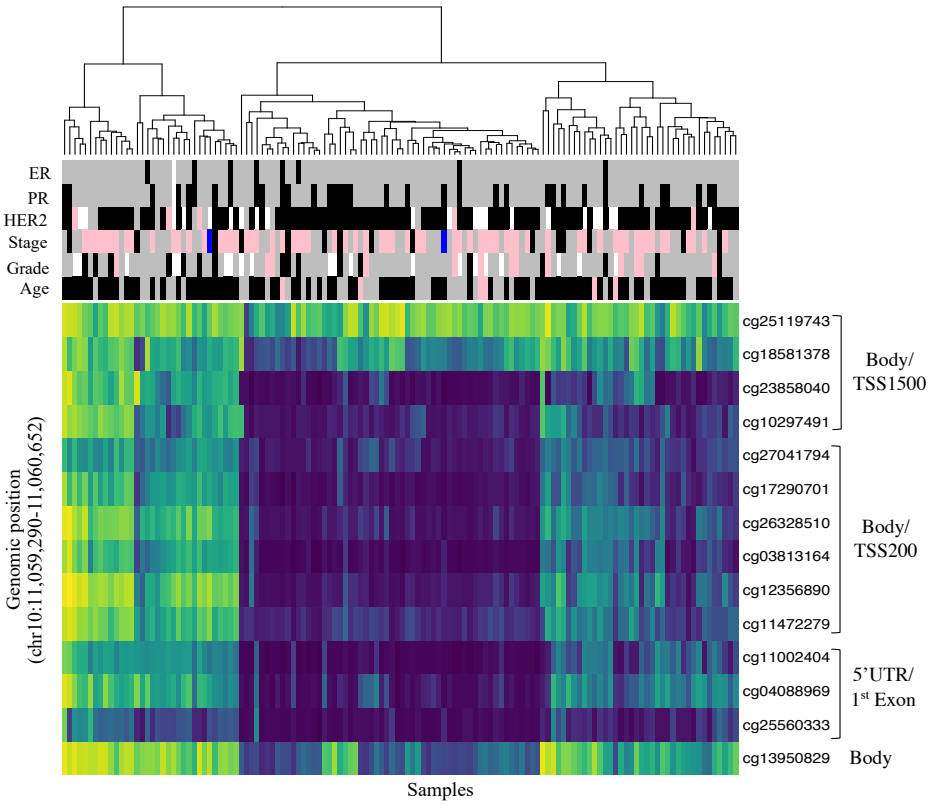
g) *HCG4P3*



h) *HES5*



i) *CELF2*



j) *EFCAB4B*

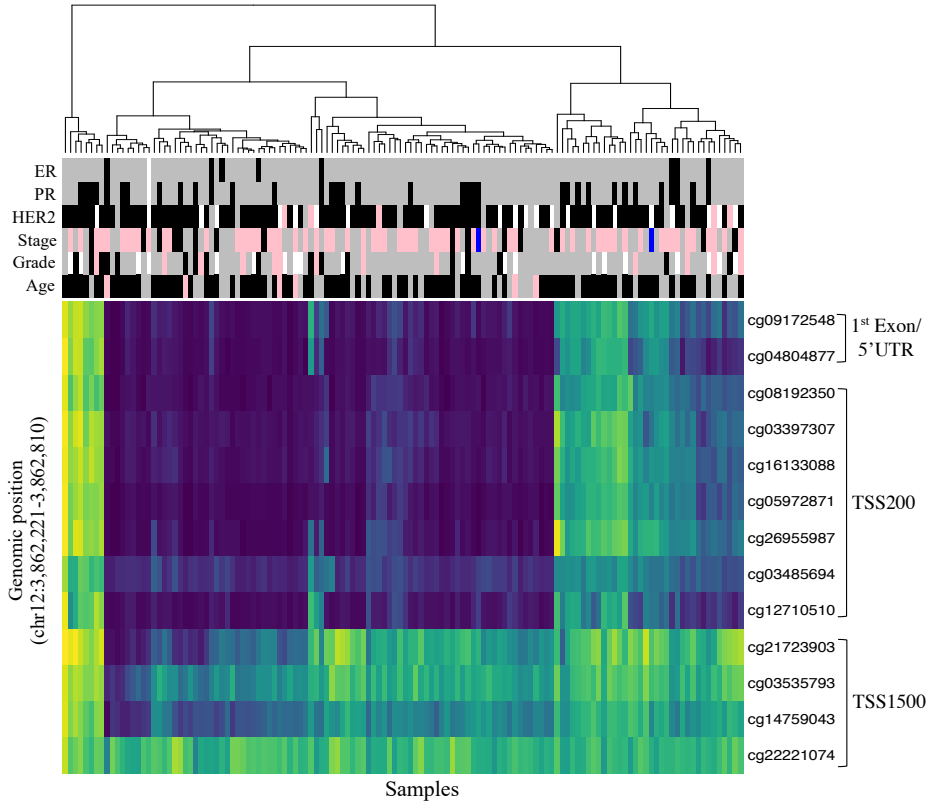


Figure 3.15: Methylation pattern of ILBC samples.

Heatmaps show the methylation patterns of invasive lobular breast cancer (ILBC) samples in the Melbourne Collaborative Cohort Study (MCCS) across the ten most significant variably methylated regions (VMRs): **a** *ISM1*, **b** *APC*, **c** *TMEM101*, **d** *ASCL2*, **e** *NKX6*, **f** *HIST3H2A*, **g** *HCG4P3*, **h** *HES5*, **i** *CELF2* **j** *EFCAB4B*. Annotation of CpGs by genomic position and location in the context of gene are marked on the maps. Annotation of samples by age at diagnosis and tumour characteristics are shown in the colour bars as indicated in the legend on the top-right. The methylation beta-value of the CpG positions shown in the heatmap is indicated in the colour key on the top-right corner.

The number of CpGs included in each VMR was wide-ranging (between 11 and 52) with some tendency for VMRs including more CpGs to be more highly ranked (Figure 3.16).

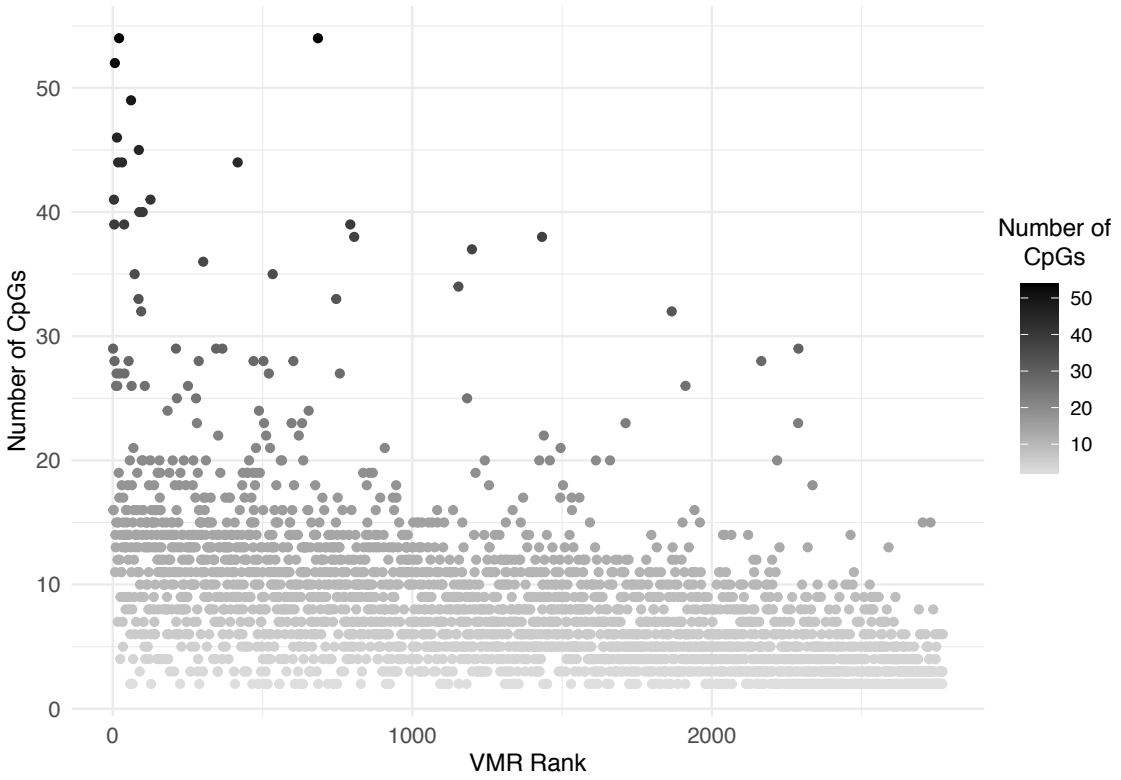


Figure 3.16: Relation between the number of CpGs related to each variably methylated region and their ranking.

The graphic shows the distribution of total number of CpG positions related to the variably methylated regions (VMRs), n=2,771, identified within ILBC samples in the Melbourne Collaborative Cohort Study (MCCS). The ranks of the VMRs are shown on the x-axis and the number of CpG positions related to each VMR is shown on the y-axis and represented by the colour gradient as indicated in the legend on the right side of the plot.

A significant enrichment for CpG-island associated regions compared to all probes included in the HM450K array was identified (Figure 3.17a). Gene annotation also showed that 62% of the VMRs were located in gene promoter regions (1st Exon, 5 prime UTR, TSS1500 and TSS200) compared with 20% in gene body regions and 23% in enhancer regions (Figure 3.17b).

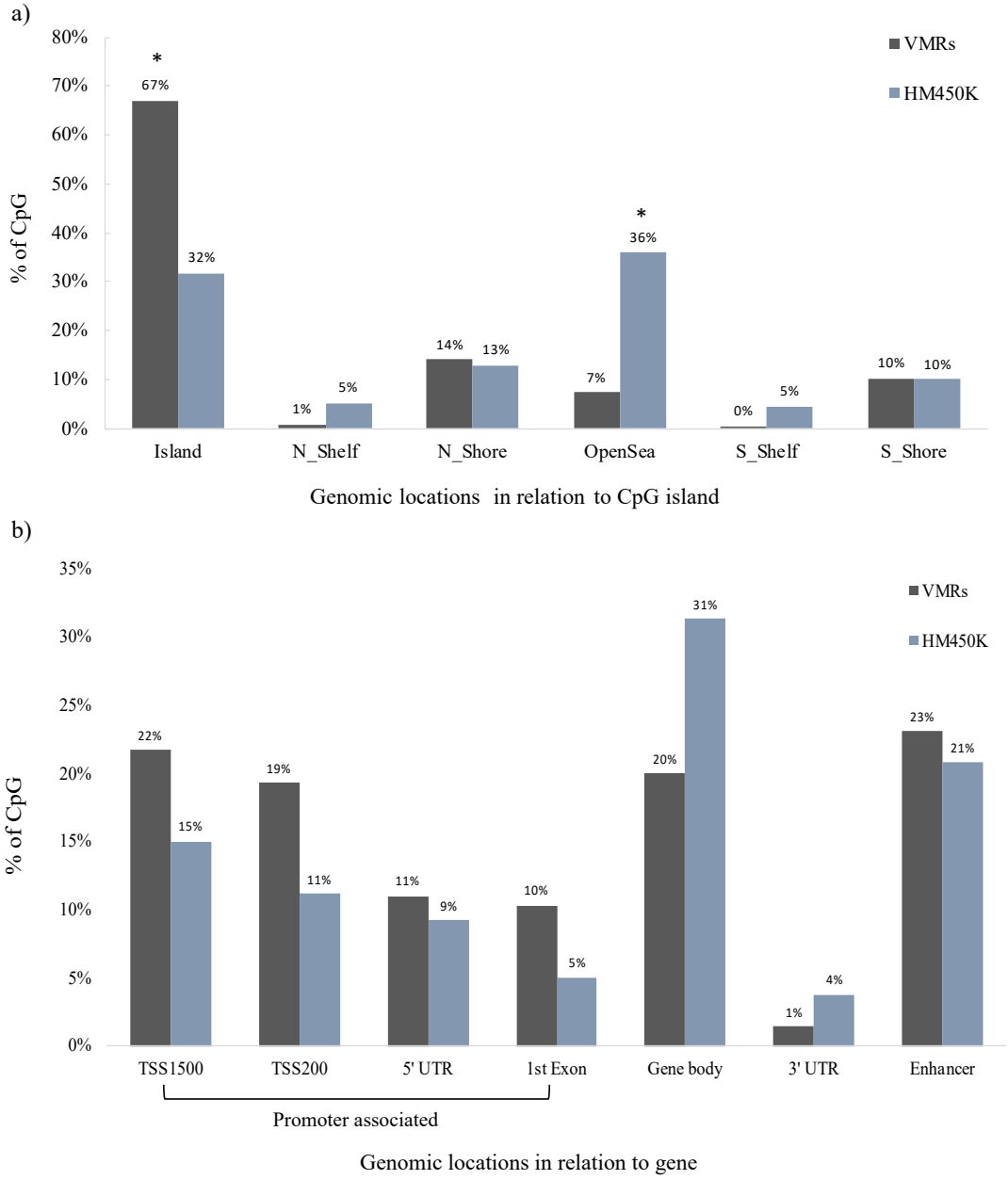


Figure 3.17: Genomic distribution of the variably methylated regions.

Bar plots show the distribution of 2,771 variably methylated regions (VMRs) identified within ILBC samples in the Melbourne Collaborative Cohort Study (MCCS) **a**) relative to CpG islands, shores (0-2 kb from island), shelves (2-4 kb from island) and open sea and **b**) in relation to the gene. Different genomic locations are shown on the x-axis and the percentage of CpG positions related to the VMRs is shown on the y-axis. The distribution of the HM450K probes relative to each CpG context is also indicated. P-values (Chi-square test) assessing significant enrichment in a given category relative to the HM450K array composition are indicated (* $P < 0.001$).

The pathway enrichment analysis showed that the genes associated with the VMRs were enriched for 1,973 terms (FDR-adjusted P <0.05) including 54 KEGG pathways with stronger evidence for *neuroactive ligand-receptor interaction* (hsa04080), *breast cancer* (hsa05224), *pathways in cancer* (hsa05200), *hippo signalling pathway* (hsa04390), *Rap1 signalling pathway* (hsa04015) and *PI3K-Akt signalling pathway* (hsa04151). Figure 3.18 shows the twenty most significant KEGG pathways enriched in the VMRs.

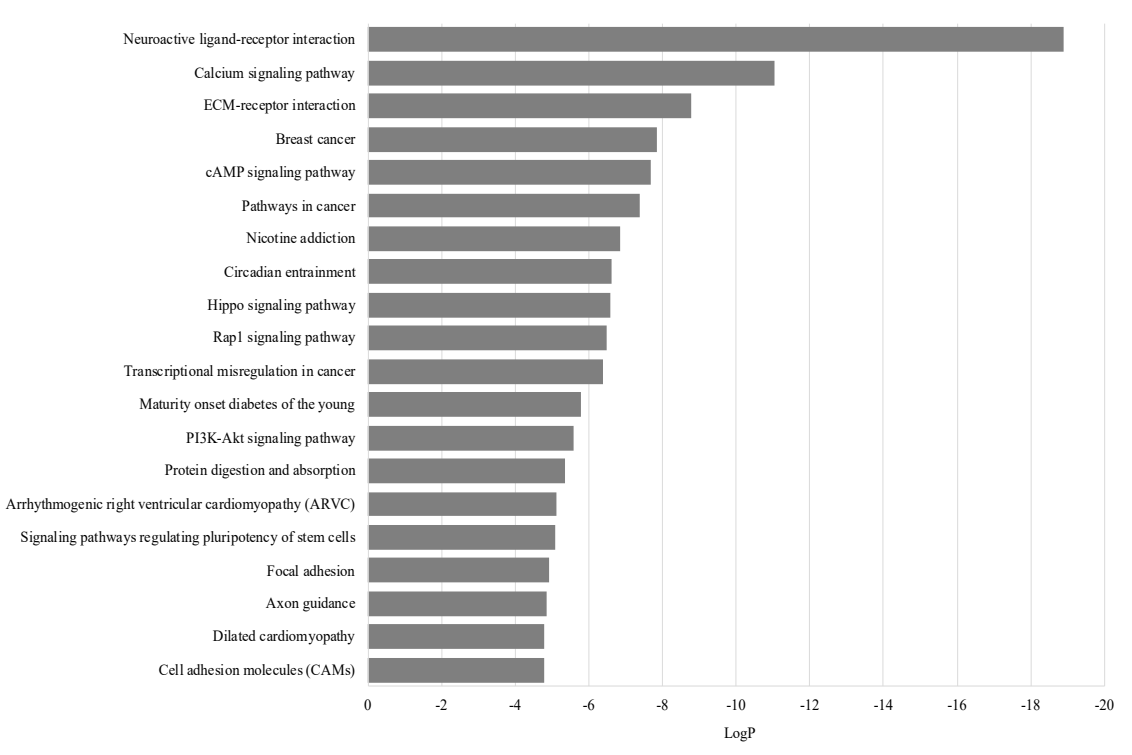


Figure 3.18: Twenty most significantly enriched KEGG pathways.

Bar plot shows twenty most significantly enriched KEGG pathways in the variably methylated region (VMRs) identified within ILBC samples in the Melbourne Collaborative Cohort Study (MCCS). The enriched terms are shown on the y-axis and the P-values (log transformed) assessing significant enrichment are shown on the x-axis.

Replication of the VMR analysis in TCGA dataset (n=168), identified 2,760 VMRs, of which 763 (28%) overlapped with the MCCS. The ten most significant VMRs identified in the MCCS ranked highly in the TCGA dataset (Table 3.4). Pathway enrichment analysis of the 763 overlapping VMRs resulted in 416 enriched functional terms (FDR-adjusted P <0.05) including nine enriched KEGG pathways. Of these, 369 overlapped with pathways identified for all MCCS VMRs; *neuroactive ligand-receptor interaction* (hsa04080) and *hippo signalling pathway* (hsa04390) were among the KEGG pathways that were also found to be significantly enriched using all MCCS VMRs.

Table 3.4: Ten most significant variably methylated regions identified in the Melbourne Collaborative Cohort Study and their respective ranking in The Cancer Genome Atlas dataset.

MCCS[a]					TCGA[b]	
Genomic location of the VMR[c] (hg19)	**minfdr***	**Number of CpGs**	**Associated gene[†]**	**Genomic location of the VMRs in relation to the corresponding genes**	**minfdr***	**Rank in TCGA**
chr20:13199787-13201844	$5x10^{-181}$	29	*ISM1*	TSS1500	$1x10^{-120}$	10
chr5:112073348-112074043	$5x10^{-181}$	16	*APC*	Body, 1st exon, TSS200, TSS1500	$3x10^{-170}$	4
chr17:42091713-42093050	$4x10^{-172}$	16	*TMEM101*	TSS1500, TSS200, 5'UTR, 1st exon	$3x10^{-92}$	20
chr11:2290953-2293552	$2x10^{-152}$	41	*ASCL2*	3'UTR, 1st exon, 5'UTR, TSS200, TSS1500	$1x10^{-90}$	23
chr10:134598496-134602228	$1x10^{-142}$	39	*NKX6*	Body, 1stExon, TSS1500	$6x10^{-118}$	12
chr1:228644750-228647248	$1x10^{-131}$	28	*HIST3H2A/ HIST3H2BB*	TSS1500, TSS200	$2x10^{-196}$	2
chr6:29973557-29976071	$4x10^{-124}$	52	*HCG4P3/HLA-J*	Body	$2x10^{-194}$	3
chr1:2460621-2462364	$1x10^{-110}$	11	*HES5*	3'UTR, Body, TSS200, TSS1500	$3x10^{-90}$	24
chr10:11059290-11060652	$2x10^{-109}$	14	*CELF2*	TSS1500, TSS200, 5'UTR, 1st exon	$9x10^{-130}$	7
chr12:3862221-3862810	$6x10^{-104}$	13	*EFCAB4B*	1stExon, 5'UTR, TSS200, TSS1500	$7x10^{-198}$	1

[a] Melbourne Collaborative Cohort Study. [b] The Cancer Genome Atlas. [c] Variably methylated region. * minimum adjusted *P*-value. [†] RefSeq gene name. TSS200 is the region from the transcript start site (TSS) to 200 nucleotides upstream of TSS; TSS1500 is the region from 200 to 1500 nucleotides upstream of TSS; 5' UTR is the region within 5 prime untranslated regions, between the TSS and the ATG start site; Body is the region between the ATG and stop codon; 3' UTR is between the stop codon and poly A signal.

3.3.15 VMRs and association with overall survival

In the MCCS, higher tumour methylation showed association with shorter overall survival for *APC* (HR = 1.28, 95% CI: 1.07-1.53), *HIST3H2A/HIST3H2BB* (HR = 1.28, 95% CI: 1.02-1.62), *CELF2* (HR = 1.30, 95% CI: 1.07-1.58) and *TMEM101* (HR = 1.21, 95% CI: 1.00-1.48). Weak evidence of association was also observed for *ISM1* (HR = 1.34, 95% CI: 0.97-1.85), *NKX6* (HR = 1.25, 95% CI: 0.98-1.60) and *HCG4P3* (HR = 1.24, 95% CI: 0.93-1.67). After adjusting for age at diagnosis and tumour stage, the association remained consistent for *APC* (HR = 1.24, 95% CI: 1.04-1.49), *TMEM101* (HR = 1.22, 95% CI: 0.99-1.51) and *HCG4P3* (HR = 1.25, 95% CI: 0.91-1.72) (Table 3.5). As shown in Table 3.5, all VMRs had an average methylation level below 0.5 and the direction of association was positive (gains in methylation associated with shorter survival).

In TCGA dataset, the crude HRs were all positive, consistent with the MCCS dataset, albeit generally greater, in particular for *ISM1* (HR = 1.48, 95% CI: 0.91-2.41), *ASCL2* (HR = 1.28, 95% CI: 0.74-2.20), *NKX6* (HR = 2.06, 95% CI: 1.32-3.21), *HIST3H2A/HIST3H2BB* (HR = 1.35, 95% CI: 1.00-1.83), *HCG4P3* (HR = 2.04, 95% CI: 1.32-3.15), *CELF2* (HR = 1.50, 95% CI: 1.06-2.12) and *EFCAB4B* (HR = 1.41, 95% CI: 1.05-1.89). Associations remained consistent after adjustment for age at diagnosis and tumour stage for all VMRs except those located at *APC* and *HES5*. The pooled HRs after adjustment for age at diagnosis and tumour stage showed that methylation was associated with overall survival for four genes: *APC* (HR = 1.18, 95% CI: 1.02-1.36), *TMEM101* (HR = 1.23, 95% CI: 1.02-1.48), *HCG4P3* (HR = 1.37, 95% CI: 1.05-1.79) and *CELF2* (HR = 1.21, 95% CI: 1.02-1.43) (Table 3.6).

Table 3.5: Hazard ratios for the association between the methylation levels at the ten most significant variably methylated regions and overall survival in the Melbourne Collaborative Cohort Study and The Cancer Genome Atlas dataset.

	MCCS[a]							TCGA[b]						
				Adjusted for age		Adjusted for age and stage					Adjusted for age		Adjusted for age and stage	
Gene[†]	**Average methy-lation***	**HR (95% CI)**	***P***	**HR (95% CI)**	***P***	**HR (95% CI)**	***P***	**Average methy-lation***	**HR (95% CI)**	***P***	**HR (95% CI)**	***P***	**HR (95% CI)**	***P***
APC	39.1	1.28 (1.07-1.53)	0.01	1.24 (1.04-1.47)	0.02	1.24 (1.04- 1.49)	0.01	35.9	1.16 (0.89-1.51)	0.28	1.12 (0.86-1.44)	0.41	1.06 (0.82- 1.38)	0.63
TMEM101	39.6	1.21 (1.00-1.48)	0.06	1.19 (0.98-1.44)	0.08	1.22 (0.99- 1.51)	0.06	39	1.13 (0.77-1.66)	0.52	1.12 (0.78-1.60)	0.54	1.27 (0.87- 1.85)	0.21
ISM1	22.8	1.34 (0.97-1.85)	0.07	1.02 (0.76-1.38)	0.89	0.90 (0.65- 1.26)	0.54	19.2	1.48 (0.91-2.41)	0.11	1.38 (0.80-2.37)	0.25	1.48 (0.86- 2.54)	0.15
ASCL2	44.4	1.16 (0.81-1.65)	0.42	0.93 (0.66-1.31)	0.69	0.99 (0.71- 1.38)	0.95	44.6	1.28 (0.74-2.20)	0.38	1.17 (0.68-2.02)	0.57	1.44 (0.81- 2.57)	0.22
HIST3H2A	20.1	1.28 (1.02-1.62)	0.03	1.08 (0.86-1.35)	0.53	1.03 (0.82- 1.29)	0.78	20.1	1.35 (1.00-1.83)	0.05	1.28 (0.94-1.73)	0.12	1.23 (0.90- 1.68)	0.18
NKX6	29	1.25 (0.98-1.60)	0.07	1.06 (0.83-1.35)	0.63	1.01 (0.79- 1.29)	0.91	30.2	2.06 (1.32-3.21)	0.001	1.88 (1.21-2.92)	0.01	2.01 (1.28- 3.17)	0.002
HCG4P3	29.1	1.24 (0.93-1.67)	0.14	1.13 (0.84-1.53)	0.41	1.25 (0.91- 1.72)	0.16	31.1	2.04 (1.32-3.15)	0.001	1.80 (1.13-2.85)	0.01	1.69 (1.05- 2.72)	0.03
HES5	19	1.11 (0.88-1.40)	0.38	1.11 (0.88-1.40)	0.37	1.13 (0.89- 1.42)	0.29	20.9	1.26 (0.88-1.80)	0.21	1.15 (0.80-1.65)	0.45	1.13 (0.76- 1.68)	0.53
CELF2	33.4	1.30 (1.07-1.58)	0.01	1.12 (0.93-1.36)	0.23	1.13 (0.93- 1.36)	0.21	35	1.50 (1.06-2.12)	0.02	1.44 (1.01-2.05)	0.04	1.51 (1.07- 2.13)	0.02
EFCAB4B	32.8	1.01 (0.83-1.23)	0.88	0.96 (0.80-1.15)	0.63	0.99 (0.83- 1.19)	0.99	34.5	1.41 (1.05-1.89)	0.02	1.32 (0.98-1.78)	0.07	1.25 (0.93- 1.67)	0.14

[a] Melbourne Collaborative Cohort Study. [b] The Cancer Genome Atlas. [†] Gene associated with the variably methylated regions (VMRs), most of the VMRs were located in the promoter region of the genes, RefSeq gene name * Average methylation level (beta-value) of the samples across the VMRs, HR: Hazard ratio, CI: Confidence interval.

Table 3.6: Pooled hazard ratios for the association between methylation levels at the ten most significant variably methylated regions and overall survival: Meta-analysis of the Melbourne Collaborative Cohort Study and The Cancer Genome Atlas results.

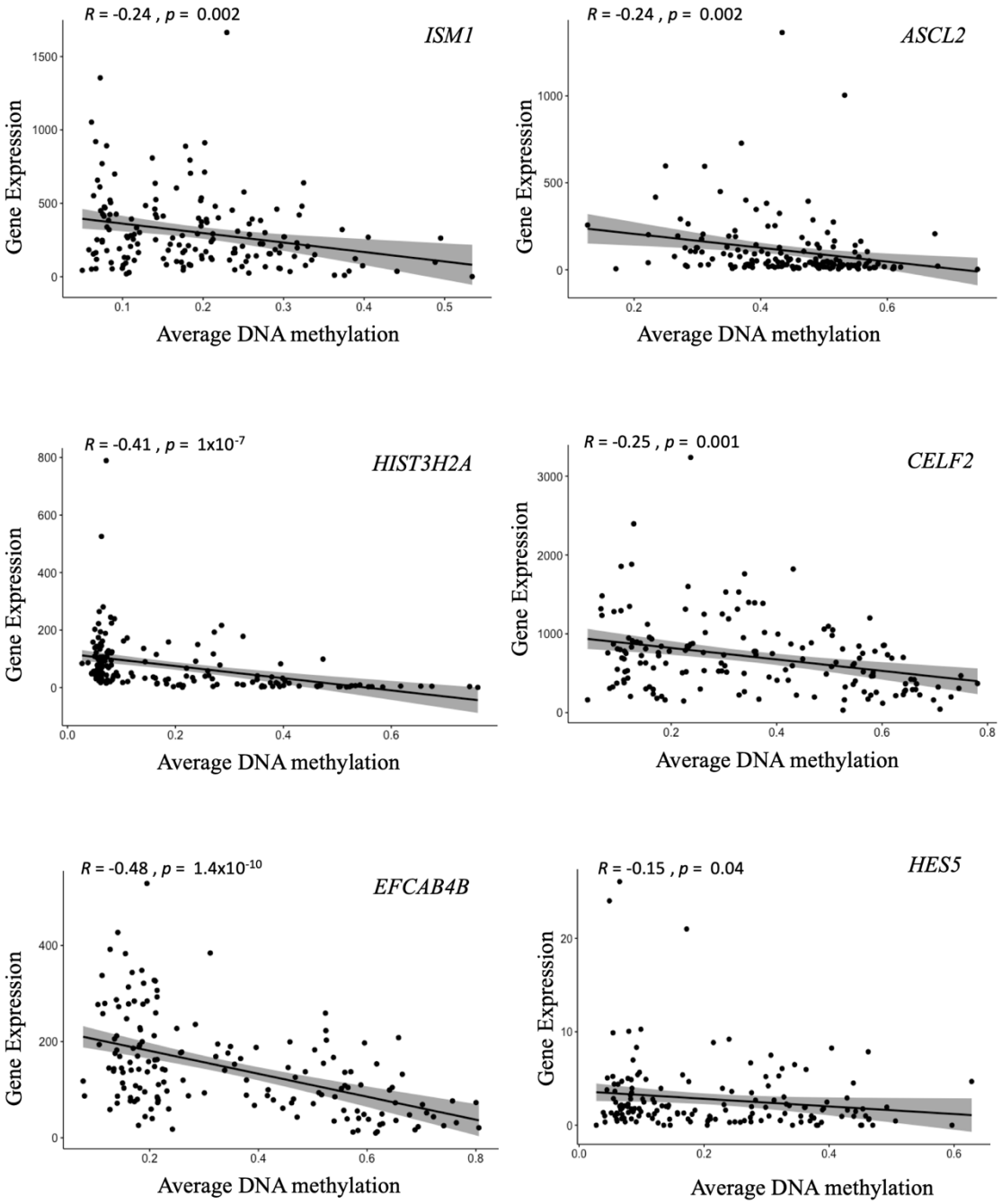
			Adjusted for age		Adjusted for age and stage	
Gene[†]	**HR (95% CI)**	***P***	**HR (95% CI)**	***P***	**HR (95% CI)**	***P***
APC	1.24 (1.07-1.44)	0.004	1.20 (1.04-1.39)	0.01	1.18 (1.02-1.36)	0.03
TMEM101	1.19 (1.00-1.42)	0.05	1.17 (0.99-1.39)	0.06	1.23 (1.02-1.48)	0.03
ISM1	1.38 (1.05-1.80)	0.02	1.09 (0.84-1.42)	0.50	1.03 (0.77-1.36)	0.83
ASCL2	1.19 (0.88-1.61)	0.24	0.99 (0.74-1.33)	0.96	1.08 (0.81-1.45)	0.57
HIST3H2A	1.30 (1.08-1.57)	0.004	1.15 (0.95-1.37)	0.14	1.09 (0.91-1.31)	0.33
NKX6	1.40 (1.13-1.74)	0.002	1.21 (0.98-1.50)	0.08	1.18 (0.95-1.46)	0.13
HCG4P3	1.45 (1.13-1.85)	0.003	1.30 (1.00-1.67)	0.04	1.37 (1.05-1.79)	0.02
HES5	1.15 (0.95-1.40)	0.15	1.12 (0.92-1.36)	0.25	1.13 (0.92-1.38)	0.23
CELF2	1.34 (1.13-1.60)	0.0006	1.18 (1.00-1.40)	0.05	1.21 (1.02-1.43)	0.02
EFCAB4B	1.12 (0.95-1.32)	0.17	1.05 (0.89-1.22)	0.57	1.05 (0.90-1.23)	0.49

[†] Gene associated with the variably methylated regions (VMRs) most of the VMRs were located in the promoter region of the genes. [†] RefSeq gene name. HR: Hazard ratio. CI: Confidence interval.

3.3.16 Correlation with gene expression

A negative correlation between DNA methylation and gene expression was observed for six of the nine tested VMRs in TCGA (Figure 3.19). These included *EFCAB4B* (R = -0.5, P-value = 1.4×10^{-10}), *CELF2* (R = -0.25, P-value = 0.001), *HIST3H2A* (R = -0.41, P-value = 1×10^{-7}), *ASCL2* (R = -0.24, P-value = 0.002), *ISM1* (R = -0.24, P-value = 0.002) and *HES5* (R = -0.15, P-value = 0.04) (Figure 3.19a). No or slightly positive correlation between DNA methylation and gene expression levels was observed for *APC*, *TMEM101* and *NKX6* (Figure 3.19b). The feature-by-feature analysis of correlations with gene expression was very consistent with the analysis using average methylation, virtually all associations being in the same direction, with only moderate variation in effect estimates.

a)


R = -0.24 , p = 0.002
ISM1
Gene Expression
Average DNA methylation
R = -0.24 , p = 0.002
ASCL2
Gene Expression
Average DNA methylation
R = -0.41 , p = 1x10-7
HIST3H2A
Gene Expression
Average DNA methylation
R = -0.25 , p = 0.001
CELF2
Gene Expression
Average DNA methylation
R = -0.48 , p = 1.4x10-10
EFCAB4B
Gene Expression
Average DNA methylation
R = -0.15 , p = 0.04
HES5
Gene Expression
Average DNA methylation

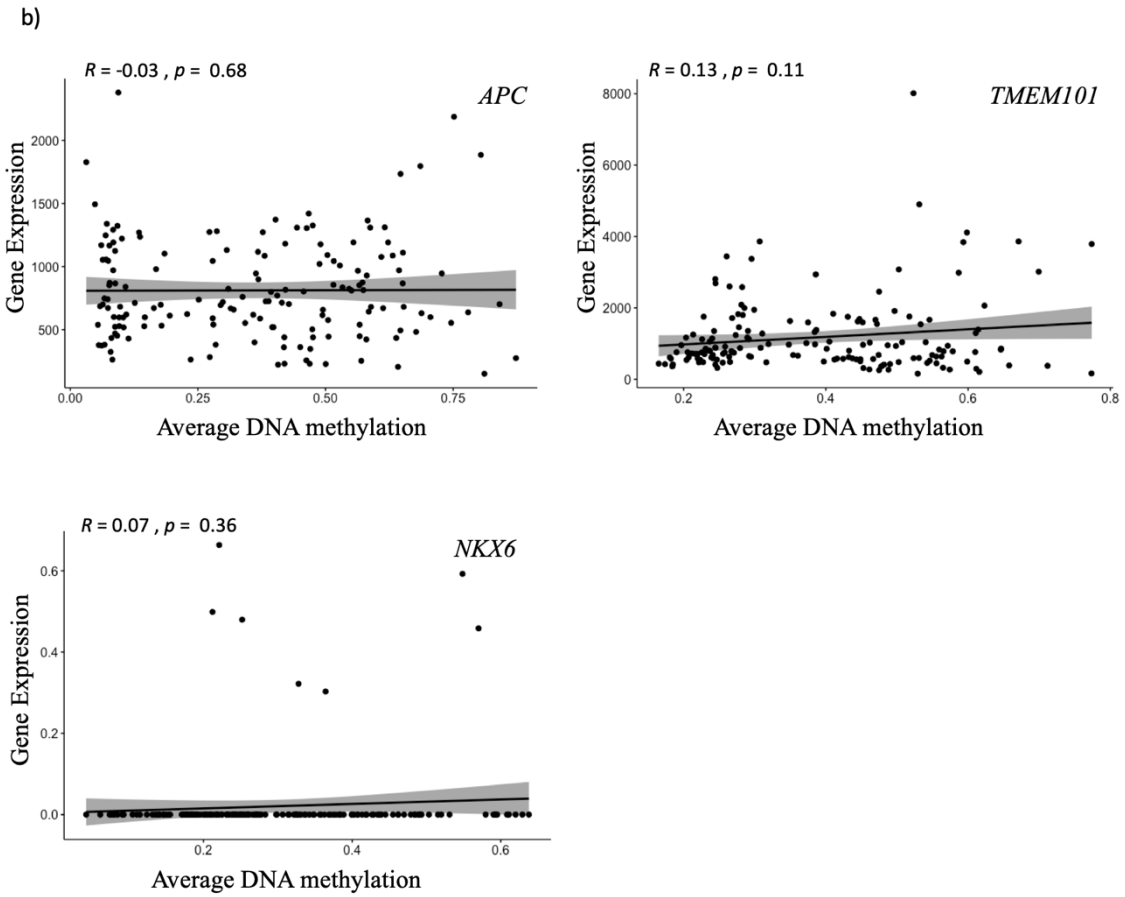


Figure 3.19: Correlation between methylation levels and gene expression.

Graphics showing the correlation between average DNA methylation (beta-value) on the x-axis and gene expression (normalised count) on the y-axis of ILBC cases at the corresponding genes associated with nine of ten strongest variably methylated regions (VMRs) with available gene expression data in The Cancer Genome Atlas (TCGA) in the ILBC cases in TCGA dataset. Corresponding gene names are marked on the top-right and the correlation value and *P*-value are marked on the top-left of the plot.

3.4 Discussion

This chapter presents the findings of the investigation of DNA methylation alterations in ILBC tumours based on a candidate gene approach and a genome-wide approach. Comparison of genome-wide DNA methylation levels of ILBC and non-ILBC tumours identified differentially methylated genes and provided some evidence that ILBC display a unique genome-wide DNA methylation profile compared with non-ILBC tumours. Methylation markers predictive of patient outcome that supported the hypothesis that tumour DNA methylation levels can be used as a prognostic biomarker for women with ILBC were also identified.

Investigating the methylation pattern of ILBC tumours using a candidate gene approach, a higher rate of promoter methylation in ILBC compared with non-ILBC tumours at *APC* (44% of ILBC *versus* 28% of non-ILBC), *RASSF1* (53% of ILBC *versus* 48% of non-ILBC), *TWIST1* (16% of ILBC *versus* 13% of non-ILBC) and *BRCA1* (64% of ILBC *versus* 58% of non-ILBC) was identified. No hypermethylation was observed at the promoter associated regions of *CDH1*, *ADAM33* and *BRCA2*, which was inconsistent with previous reports suggesting a frequent *ADAM33* promoter methylation in ILBC compared with non-ILBC tumours (Seniski *et al.*, 2009). *CDH1* promoter methylation has also been reported as an alternative method for *CDH1* silencing reported in 40-80% of ILBC tumours (Droufakou *et al.*, 2001; Sarrió *et al.*, 2003; Caldeira *et al.*, 2006). However, no hypermethylation was observed at *CDH1* promoter in this study. One of the reasons for this discrepancy could be the difference in technology. While Illumina HM450K assay was used to in this study, all the studies mentioned above have used methylation-specific PCR for measuring the methylation levels.

A distinct methylation pattern was observed across the two *APC* promoters, promoter 1A and promoter 1B. While promoter 1B showed hypomethylation across all the ILBC tumours, a substantial variability in methylation level was observed across promoter 1A. *APC* promoter 1A methylation has previously been reported in breast cancer. Liu *et al.*, (2007) reported promoter 1A hypermethylation in 28/76 (37%) of breast cancer cases compared with normal breast tissue (0%) (*P*-value < 0.05). They also reported a negative correlation between promoter 1A hypermethylation and APC protein expression (measured by immunohistochemistry) (*R*= -0.368, *P*-value < 0.05), suggesting a functional role of promoter 1A methylation in *APC* gene regulation (Liu *et al.*, 2007). A higher frequency of *APC* promoter 1A methylation in ILBC tumours compared with non-ILBC and a lack of methylation in the normal adjacent breast samples indicates its potential applicability in early diagnosis of ILBC and should further be investigated.

Studying the methylation pattern at individual genes revealed that different genomic regions associated with the gene show distinct pattern of methylation. While methylation level across the gene body region was found to be higher at all the genes, the different genomic regions associated with the gene promoter showed varied methylation patterns. For instance, within the promoter associated region of *BRCA1*, the 5 prime UTR region was hypomethylated across all the samples except one, whereas the TSS1500 region was hypermethylated in 96/151 (64%) of ILBC tumours (Figure 3.9). This may suggest that different regions associated with a gene promoter may have different functional control in gene expression regulation. This has previously been suggested in a study that reported that methylation level at the 1st exon is tightly linked to the gene expression (Brenet *et al.*, 2011). The one sample hypermethylated at the 5 prime UTR of *BRCA1* was grade III tumour of mixed lobular-ductal morphology that was found to be

negative for ER expression. This may suggest that methylation at 5 prime UTR could be an indicator of an aggressive tumour behaviour and future study investigating the association between 5 prime UTR methylation in relation to *BRCA1* gene expression could unravel any possible link. This analysis again points towards the heterogeneous nature of ILBC tumours and reiterates the importance of investigating the histological subtypes for a precise understanding of their molecular nature. Future studies investigating the methylation profile of different well-defined histological subtypes of ILBC at genome-wide scale or at specific breast cancer genes will shed more light on this and would suggest approaches of using methylation level information for treating ILBC with precision.

Comparing the genome-wide DNA methylation patterns of ILBC and non-ILBC tumours, revealed regions of differential methylation between the two tumour types. The DMRs were significantly enriched in pathways such as *metabolism of RNA* (R-HSA-8953854), *mRNA processing* (GO:0006397), *RNA splicing* (GO:0008380), *cell cycle* (R-HSA-1640170) and *DNA repair* (GO:0006281). Many of the genes involved in these pathways have been found to be dysregulated in cancer suggesting the involvement of DMRs in cancer initiation and progression. Limiting the analysis to only luminal A ILBC and luminal A non-ILBC tumours, a similar differential methylation pattern was observed and the most significant DMRs in ILBC *versus* non-ILBC comparison remained significant in luminal A ILBC *versus* luminal A non-ILBC comparison also.

The genome-wide DNA methylation pattern of ILBC tumours, with the aim of identifying methylation markers predictive of patient outcome was investigated. Scanning of the ILBC methylome revealed regions of variable methylation in ILBC tumours. The

VMRs were primarily located in CpG island regions and were significantly enriched in pathways such as *breast cancer* (hsa05224), *pathways in cancer* (hsa05200), *hippo signalling pathway* (hsa04390), *Rap1 signalling pathway* (hsa04015) and *PI3K-Akt signalling pathway* (hsa04151). These pathways have previously been found to be dysregulated in cancer tissue suggesting the involvement of the identified VMRs in cancer initiation and progression (Bailey *et al.*, 2009; Hall *et al.*, 2010; McSherry *et al.*, 2011; Yin & Zhang, 2011; N Li *et al.*, 2017; X-L Ma *et al.*, 2019). Some of the key genes involved in the enriched pathways included *APC*, *DAPK1, BMP2 and CCND2. DAPK1* is an important regulator of cell apoptotic pathways (Gozuacik *et al.*, 2008) and *DAPK1* promoter hypermethylation has previously been reported in ILBCs with a potential role in tumour progression (Lehmann *et al.*, 2002; Tserga *et al.*, 2012). *BMP2* is a member of the TGF-ß superfamily and is involved in cell proliferation and differentiation during tumour formation (Thawani *et al.*, 2010). Promoter methylation of *BMP2* has been associated with breast cancer progression and drug resistance (M Du *et al.*, 2014). *CCND2* promoter methylation was previously reported to be a common event in breast cancer and have prognostic value (Hung *et al.*, 2018). A similar DNA methylation variability profile was observed in TCGA dataset, in particular for the VMRs showing strongest variability in the MCCS. Several previous studies have reported tumour DNA methylation to have prognostic value in cancer (Buffart *et al.*, 2008; Ellinger *et al.*, 2008; CY Hu *et al.*, 2014; Sailer *et al.*, 2017; Guo *et al.*, 2018; de Almeida *et al.*, 2019). Methylation at many gene promoters has been reported to have independent prognostic value in breast cancer including *HOXA11* (Xia *et al.*, 2017), *ESR1* and *PITX2* (Sheng *et al.*, 2017)*, HOXD13* (Zhong *et al.*, 2015) *CDH22* (Martín-Sánchez *et al.*, 2017) *BRCA1* and *RASSF1* (Jiang *et al.*, 2012; Wu *et al.*, 2013). Tumour DNA methylation and its prognostic significance has also been investigated for certain breast cancer subtypes, in

particular gene expression-based subtypes. Thomas et al., (2017) used hierarchical clustering based on DNA methylation to further segregate luminal A tumours into two subgroups and found that the subgroup with lower relative methylation showed better prognosis (Fleischer *et al.*, 2017), similar to the findings of this study. Another study using whole-genome methylation sequencing stratified triple-negative breast cancers into three methylation-defined clusters and found the hypomethylated cluster to show better prognosis compared with the other two highly methylated clusters (Stirzaker *et al.*, 2015), also consistent with results of this study. However, to our knowledge, no study has reported on the overall tumour methylation variability in ILBC and tested the potential for the variably methylated regions to be used as prognostic markers. The assessment of VMRs was genome-scale but only the highest ranking VMRs were tested for their association with survival. Although many of the tested VMRs showed a significant association with overall survival, there could be other VMRs or individual CpG sites for which methylation is associated with survival. Promoter hypermethylation at *APC*, *TMEM101* and *HCG4P3* was found to be associated with shorter overall survival in the MCCS after adjustment for age and tumour stage. The results in TCGA were largely consistent with the MCCS, although associations generally appeared stronger; this might suggest that the prognostic value of these DNA methylation markers is greater for women with more advanced ILBC. In the pooled analysis, DNA methylation at four genes (*APC, TME101, HCG4P3* and *CELF2*) was associated with shorter overall survival. All the highest ranking VMRs had an average methylation level below 0.5 and the direction of association with survival was virtually always positive, which indicates that methylation gains (i.e., loss of the normal hypomethylation state) were associated with worse survival. While a low correlation between methylation and gene expression was observed for *APC*, *TMEM101* and *NKX6,* six of the nine VMRs tested showed a strong inverse correlation.

This could suggest that the VMRs tend to overlap with the promoter regions for the corresponding genes and the hypermethylation may be involved in repression of these genes. *APC* is a well-known tumour suppressor gene, and this finding is in agreement with previous reports (K He *et al.*, 2016; Debouki-Joudi *et al.*, 2017). Debouki et al., (2017) found a significant correlation between *APC* promoter methylation and aggressive behaviour of both non-familial and familial breast cancer in the Tunisian population (Debouki-Joudi *et al.*, 2017). The association of *APC* promoter methylation with reduced survival has also been reported for other cancer types, such as non-small cell lung cancer (Brabender *et al.*, 2001) and prostate cancer (Henrique *et al.*, 2007; Richiardi *et al.*, 2009). CELF2*,* an RNA binding protein involved in alternative splicing, has also been reported to be involved in breast cancer growth and progression. Piqué *et al.,* (2019) found that *CELF2* promoter methylation led to a loss of CELF2 expression that had a growth promoter effect in breast tumours. They also found that *CELF2* promoter methylation was associated with worse patient outcome (Piqué *et al.*, 2019). In TCGA data, a strong, negative correlation between *CELF2* promoter methylation and the gene expression levels was found. TMEM101 is a transmembrane protein that has been shown to activate NF-kappa-beta signalling pathways. There is to our knowledge no previous literature suggesting a role of *TMEM101* promoter methylation in relation to cancer progression/survival. *HCG4P3* is also known as HLA complex group 4 pseudogene 3 and there is to our knowledge no record of this gene being involved in cancer.

One of the main limitations of the candidate gene and genome-wide comparison of ILBC tumours with non-ILBC and adjacent normal samples was the selection of control samples. Although all are breast tissues, the cellular composition of the non-ILBC, ILBC and adjacent normal tissue is likely to be somewhat different and therefore likely to have

different methylation profiles. The most ideal control sample for identifying DNA methylation specific to ILBC will one that had the same types of cells as the ILBC samples, which is challenging to achieve given the variation in normal breast and breast cancer histopathological features. Another factor that may have an impact on the results of ILBC and non-ILBC comparison at the genome-wide methylation level is tumour purity. Although more than 80% of the samples in both ILBC and non-ILBC tumour sample groups showed a tumour purity of more than 50%, the samples showed a range of purity values that could have impacted the DNA methylation signals for samples with low tumour purity. No significant bias (*P*.value, t.test= 0.13) was observed between ILBC and non-ILBC samples in terms of tumour purity (Table 2.1). The main limitation of the VMR study was the relatively small sample size that limited the analysis to all-cause death as an endpoint. The MCCS and TCGA data had different characteristics in terms of their study design and sample characteristics. The two studies had different follow-up times and TCGA data had more young women and generally higher tumour stage (Table 3.3). These differences in the studies could in part account for the low concordance (28%) between the VMRs identified between the MCCS and TCGA. The findings for both the VMR and survival analysis were nevertheless consistent across the two studies. The main factors that we thought could impact methylation profiles in tumours and ILBC survival were considered, i.e., age and stage. Factors such as smoking, alcohol consumption or diabetes, and perhaps family history (via underlying genetic sequence) likely play some role, but it is presumably less important, so were not included in the analysis. These variables are not systematically collected with precision (questionnaires) in the clinical setting. In this context, this study identified methylation biomarkers and it is likely that many factors worthy of investigation (genetic and lifestyle and environmental) play a role in explaining the observed associations. Finally, while a large number of regions across

the ILBC genome that showed substantial variable methylation pattern were identified, only the strongest ten VMRs were tested for association with survival to minimise the multiple testing burden. If replicated by other studies, the methylation markers identified in our study may contribute to the development of molecular signatures for enhanced prediction of ILBC survival. Using TCGA dataset for validation, we were able to show that the VMRs identified in the MCCS were also consistently variably methylated across TCGA dataset. Although HRs in the two datasets show some variation, (including their statistical significance), they are consistent in terms of the direction of the association. Differences between the two studies, MCCS and TCGA in terms of study design and clinical characteristics of participants may in part account for the observed variation in association. This work warrants further validation in a larger dataset with less sample variation.

3.5 Summary

The analyses involving candidate gene approach in Part I of this study indicated that ILBC tumours were more frequently methylated at *APC*, *RASSF1*, *TWIST1* and across the TS1500 region of *BRCA1* compared with the non-ILBC tumours. Based on the genome-wide comparison of ILBC and non-ILBC tumours, genes that differed in their methylation patterns between ILBC and non-ILBC tumours were identified. Many of these genes were found to be enriched for pathways related to mRNA processing. Further investigation is required to find out the effect of these methylation alterations on the expression level of corresponding genes. This study also indicated that methylation levels at the most variable regions across the genome may explain differences in tumour prognosis within the ILBC subtype. *APC*, *TMEM101*, *HCG4P3* and *CELF2* promoter methylation were identified as possibly relevant prognostic biomarkers for women with

ILBC. Further studies are required to confirm these findings and to assess their utility in a clinical setting.

Chapter 4 Sub-classifying Invasive Lobular Breast Cancer Based on Tumour DNA Methylation Profiling

4.1 Introduction

Heterogeneity within a breast cancer subtype is regarded as a major challenge in personalised cancer medicine. ILBC also presents as a heterogeneous disease showing varying morphological features, genetic makeup, clinical presentation and patient outcome (section 1.5) of the thesis. However, they are routinely regarded as a single entity while making treatment decisions.

Gene expression profile-based subtyping has highlighted the heterogeneity of breast cancer and has enabled further subtyping. In their analysis, Sørlie et al., (2001) included 78 breast tumours, out of which only five were ILBC tumours (Sørlie *et al.*, 2001). Out of the five ILBCs, two were classified as luminal A and one each as luminal B, basal and normal-like subtype. West et al., (2001) used the expression levels of 7,129 genes and identified similar subgroups as the intrinsic subtypes in 49 breast cancer samples that were all IDBC (West *et al.*, 2001). In another study, Van't Veer et al., (2002), used gene expression profiling of 98 primary breast cancer samples and identified a 70 gene signature for poor prognosis however, in this study the histological information of samples was not reported (Van't Veer *et al.*, 2002). ILBC was poorly represented in the foundation work that subtyped breast cancer and a better stratification is required to identify subtype-specific diagnostic and prognostic markers for a more precise treatment regime for ILBC.

DNA methylation is known to be a crucial regulator of gene expression and tumour methylation profiling has been shown to be a robust tool to accurately identify disease-specific subtypes in many cancer types including breast cancer (Toyota *et al.*, 1999; Noushmehr *et al.*, 2010; Barreau *et al.*, 2013; Conway *et al.*, 2014). DNA methylation profiling is less technically challenging compared with gene-expression profiling as DNA is more stable than RNA and DNA methylation is detectable in tumour tissue thus making it a more applicable tool for clinical purposes (Kit *et al.*, 2012).

Several studies have used DNA methylation levels to further stratify breast cancer. However, these studies have primarily been focused on the intrinsic subtypes or more aggressive breast cancer subtype, TNBC. DiNome et al., (2019) used genome-wide DNA methylation profiles and identified four epitypes within TNBC that showed differences in survival and their gene expression and mutation profiles (DiNome *et al.*, 2019). Zhang et al., (2018) identified nine subgroups within the intrinsic subtypes including two further subtypes within the basal-like group with prognostic significance (S Zhang *et al.*, 2018). In another study, Fleischer et al., (2017) used hierarchical clustering based on tumour DNA methylation levels to further segregate luminal A tumours into two subgroups and found that the subgroup with lower relative methylation showed better prognosis (Fleischer *et al.*, 2017). Stefansson et. al., (2015) using DNA methylation profiling of 40 tumours and 17 normal breast samples defined two DNA methylation-defined subtypes; Epi-LumB and Epi-Basal associated with poor outcome (Stefansson *et al.*, 2015).

No study has yet attempted to subtype ILBC based on tumour DNA methylation profiling. In Chapter 3, it was demonstrated that a substantial variability in tumour DNA methylation exists within ILBC and some of this variability is associated with prognosis.

This finding points towards the potential of using tumour DNA methylation to stratify ILBC into further subgroups. In this study, we hypothesised that subgroups of ILBC may be identified using genome-wide tumour DNA methylation data. We aimed to identify subgroups of ILBC via unsupervised cluster analysis using genome-wide tumour DNA methylation profiling of 151 ILBC and 341 non-ILBCs.

4.2 Method overview

4.2.1 Study participants and data

Analyses in this chapter included 492 invasive breast cancer samples. Details of the study resources are provided in Table 4.1.

Table 4.1: Study participants and data.

Study	**Breast cancer type**	
	ILBC	**Non-ILBC**
MCCS[a] (n=471)	130	341
kConFab[b] (n=6)	6	0
ABCFR[c] (n=15)	15	0
Details of the study design is presented in the Methods section 2.1.		
Total	151	341
Sample type	FFPE tumour enriched DNA Details of sample preparation from FFPE is presented in the Method section 2.3.1.	
Data information	Genome-wide DNA methylation using Illumina HM450K array (Details of the methylation assay is presented in the Methods section 2.5).	

ILBC: Invasive lobular breast cancer. Non-ILBC: Non-lobular invasive breast cancer. [a] Melbourne Collaborative cohort Study. [b] The Kathleen Cuningham Foundation Consortium for Research into Familial Breast Cancer. [c] Australian Breast Cancer Family Registry. FFPE: Formalin-fixed paraffin-embedded. HM450K: Illumina HumanMethylation 450K array.

4.3 Results

4.3.1 Unsupervised cluster analysis

To further subclassify ILBC, an unsupervised cluster analysis was performed (section 2.9.5), using the genome-wide tumour DNA methylation levels (M-values) of all ILBC (n=151) and non-ILBC breast cancer cases (n=341).

The clustering divided the breast cancer samples into two main groups; group A and group B, based on the overall similarity in their DNA methylation patterns. The main groups further separated into many small groups (Figure 4.1). The pattern of clustering and length of branches in the dendrogram shows the relatedness of the samples by their genome-wide DNA methylation profiles (Figure 4.1). Group A contained 115/151 (76%) of the ILBC samples, whereas a smaller proportion, 36/151 (31%) of ILBCs clustered into group B. ILBC samples in group A formed close clusters (samples were in close proximity), whereas this was not observed in group B (Figure 4.1). Based on the clustering of ILBC samples in group A, subgroups of ILBC were defined such that, samples that stemmed from the same branch and formed a close cluster were assigned to the same subgroup. Based on this criterion, three main subgroups of ILBC were defined; Subgroup 1 (n = 28), Subgroup 2 (n = 27) and Subgroup 3 (n = 21), shown in the boxes in the dendrogram (Figure 4.1).

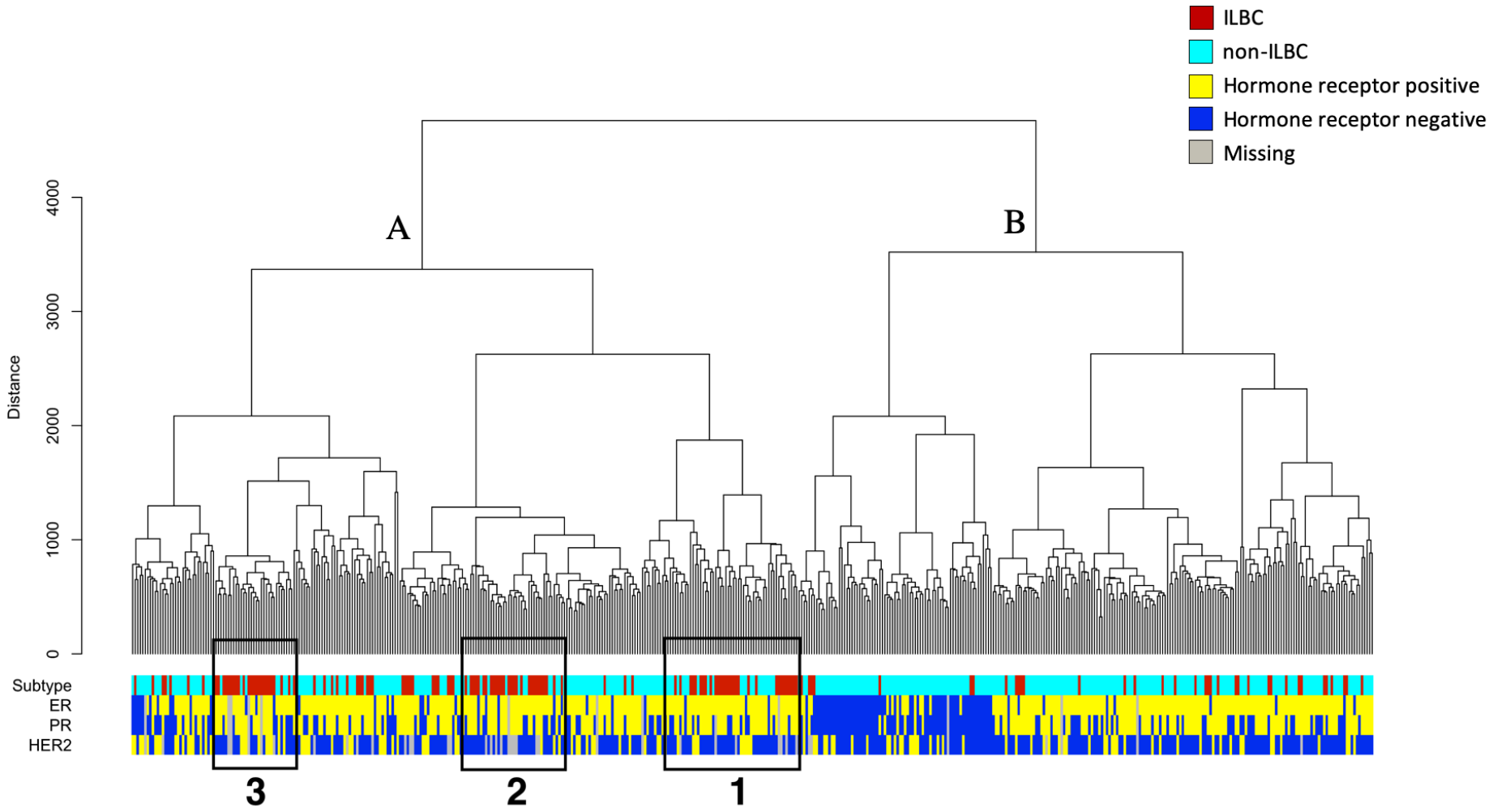

ILBC
non-ILBC
Hormone receptor positive
Hormone receptor negative
Missing
A
B
Distance
0
1000
2000
3000
4000
Subtype
ER
PR
HER2
3
2
1

Figure 4.1: Unsupervised cluster analysis of breast cancer samples (ILBC, n=151 and non-ILBC, n=341) based on their genome-wide DNA methylation profiles.

Dendrogram showing the unsupervised clustering of all breast cancer samples (n = 492) including ILBC, n = 151 and non-ILBC, n = 341, based on their genome-wide DNA methylation profiles. Each leaf of the dendrogram represents a breast cancer sample and the length of the branches show the Euclidean distance between the two clusters (y-axis). Higher distance on the dendrogram represents more dissimilar clusters and vice-versa. The colour bar "Subtype" indicates the two breast cancer histological subtypes; i.e., ILBC (shown in red) and non-ILBC (shown in turquoise). The colour bars "ER", "PR" and "HER2" indicate the estrogen, progesterone and human epidermal growth factor receptor 2 expression status, respectively of the breast tumours. Hormone receptor positive tumours are shown in "yellow" and hormone receptor negative tumours are shown in "blue" colour. The three ILBC subgroups defined based on the clustering are numbered and marked in black boxes.

As hormone receptor expression status has been known to significantly influence the DNA methylation patterns in breast tumours (Holm *et al.*, 2010), we further tested whether the clustering was driven by the ER, PR and HER2 expression status (as measured by immunohistochemistry) of the tumours. The majority of the samples in all three ILBC subgroups were ER and PR positive and HER2 negative and a strong clustering based on the hormone receptor expression status was not observed for the ILBC samples (Figure 4.1). As the breast cancer samples included in this analysis were sourced from three different studies (Table 4.1), any influence of the variation in study design on the clustering was also tested. No significant association was observed between the clustering and the studies from which the samples were sourced (chi-square test, *P*-value = 0.28).

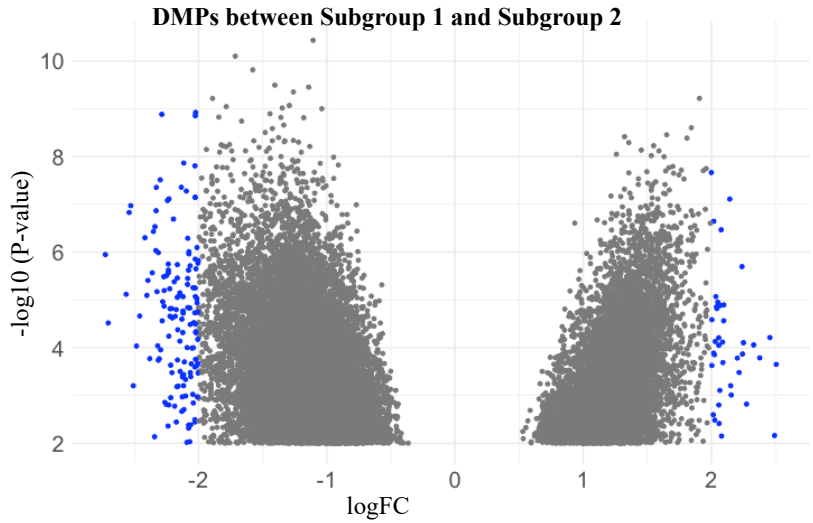
4.3.2 Differential methylation between the ILBC subgroups

To further investigate the differences in DNA methylation between the ILBC methylation-defined subgroups and to find the CpG positions that were differentially methylated (hypermethylated or hypomethylated) between the subgroups, a differential methylation analysis was performed as described in section 2.9.1 of the thesis.

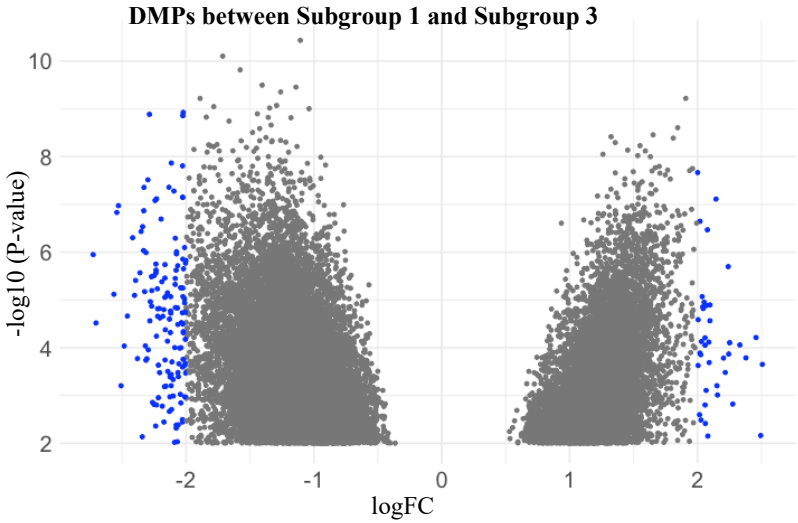
Between Subgroup 1 and Subgroup 2, 27,675 significantly (*P*-value < 0.01) differentially methylated positions (DMPs) were identified. Of these, 8,647 (31%) were hypermethylated and 19,028 (69%) were hypomethylated in Subgroup 1 compared with Subgroup 2 (Figure 4.2a). Between Subgroup 1 and Subgroup 3, 13,067 DMPs were identified, of which 4,758 (36%) were hypermethylated and 8,309 (64%) were hypomethylated in Subgroup 1 compared with Subgroup 3 (Figure 4.2b). Between

Subgroup 2 and Subgroup 3, 65 DMPs were identified and all 65 DMPs were hypomethylated in Subgroup 2 compared with Subgroup 3 (Figure 4.2c).

a)


DMPs between Subgroup 1 and Subgroup 2
-log10 (P-value)
logFC
10
8
6
4
2
-2
-1
0
1
2

b)


DMPs between Subgroup 1 and Subgroup 3
-log10 (P-value)
logFC
10
8
6
4
2
-2
-1
0
1
2

c)

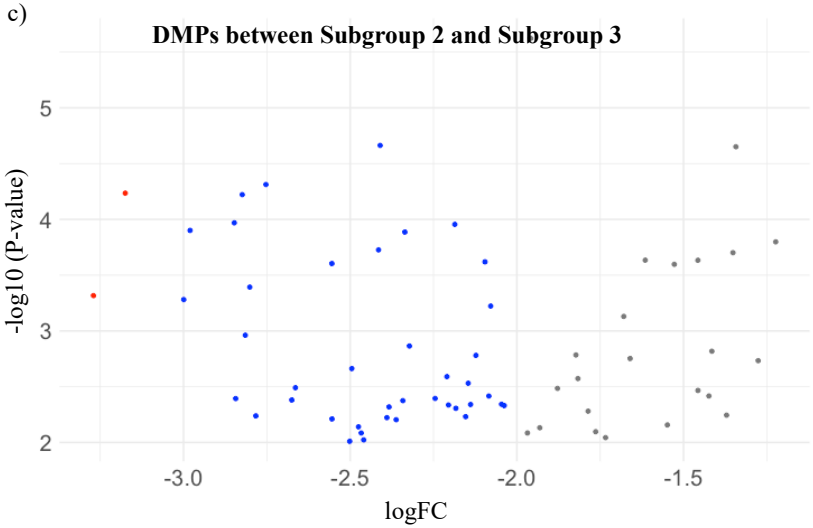

DMPs between Subgroup 2 and Subgroup 3
-log10 (P-value)
logFC
5
4
3
2
-3.0
-2.5
-2.0
-1.5

Figure 4.2: Differentially methylated positions between ILBC methylation-defined subgroups.

Volcano plots showing the differentially methylated positions (DMPs) between the ILBC methylation-defined subgroups **a**) between Subgroup 1 and Subgroup 2; **b**) between Subgroup 1 and Subgroup 3 and **c**) between Subgroup 2 and Subgroup 3. logFC (change in the average M value between the comparison subgroups) is shown on the x-axis and *P*-value (log transformed) of the DMPs is shown on the y-axis. DMPs between the comparison subgroups are represented by the dots where blue colour dots represent the DMPs with an absolute logFC of greater than 2 and the grey colour dots represent the DMPs with an absolute logFC of less than 2 between the comparison subgroups.

The differential methylation analysis revealed that Subgroup 1 was hypomethylation at most of the DMPs, whereas Subgroup 3 was hypermethylation at most of the DMPs identified between these two subgroups. The average methylation level of Subgroup 1 across the DMPs was found to be significantly different from the remaining two subgroups (Subgroup 1 and Subgroup 2- ANOVA, *P*-value = $2.7x10^{-13}$, Subgroup 1 and Subgroup 3- ANOVA, *P*-value = $7.1x10^{-12}$), whereas the methylation profiles of Subgroup 2 and Subgroup 3 were more similar (ANOVA, *P*-value = 0.33). Figure 4.3 shows the difference in average methylation levels between the ILBC methylation-defined subgroups across the DMPs between the subgroups (Figure 4.3a) and across all CpG positions genome-wide (Figure 4.3b).

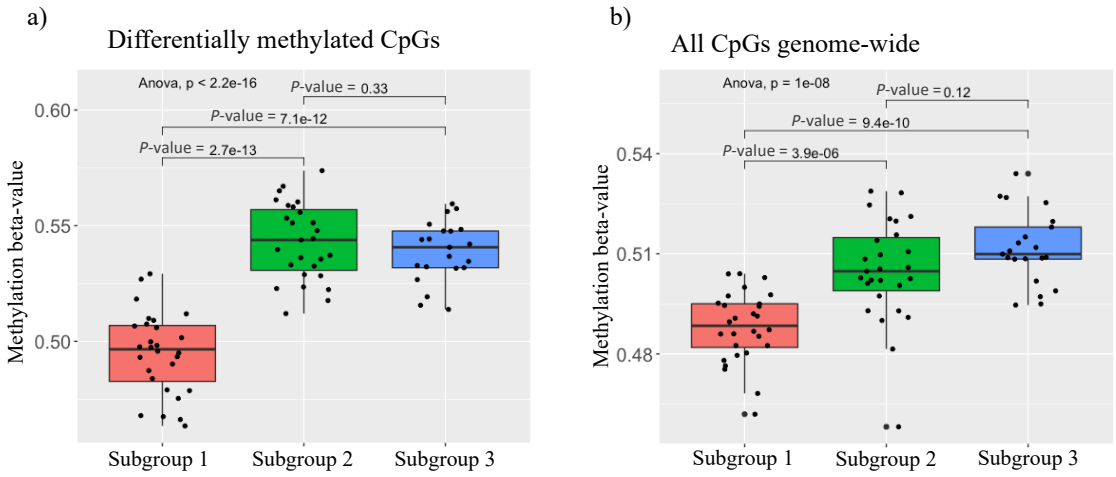


Figure 4.3: Average methylation levels (beta-value) of the ILBC methylation-defined subgroups.

Box plots illustrating the difference in average methylation levels (beta-value) between the ILBC methylation-defined subgroups; Subgroup 1, Subgroup 2 and Subgroup 3 (shown on the x-axis) across **a**) the differentially methylated positions (DMPs) between the subgroups and **b**) all CpG positions genome-wide. ANOVA, *P*-values indicating the significance of the difference in average methylation levels between the subgroups in different comparisons and the overall comparison of the three subgroups are marked on the plots.

Subgroup 1 was found to cluster alongside the TNBC cases on the dendrogram. Although, the TNBC cases were clustered into a separate branch (group B) than Subgroup 1 (Figure 4.1), the global methylation profile (average methylation beta-value across all CpG positions genome-wide) of Subgroup 1 was found to be more similar to TNBC (ANOVA, *P*-value = 0.37) than to Subgroup 2 (ANOVA, *P*-value = $3.9x10^{-6}$) and Subgroup 3 (ANOVA, *P*-value = $9.4x10^{-10}$) (Figure 4.4).

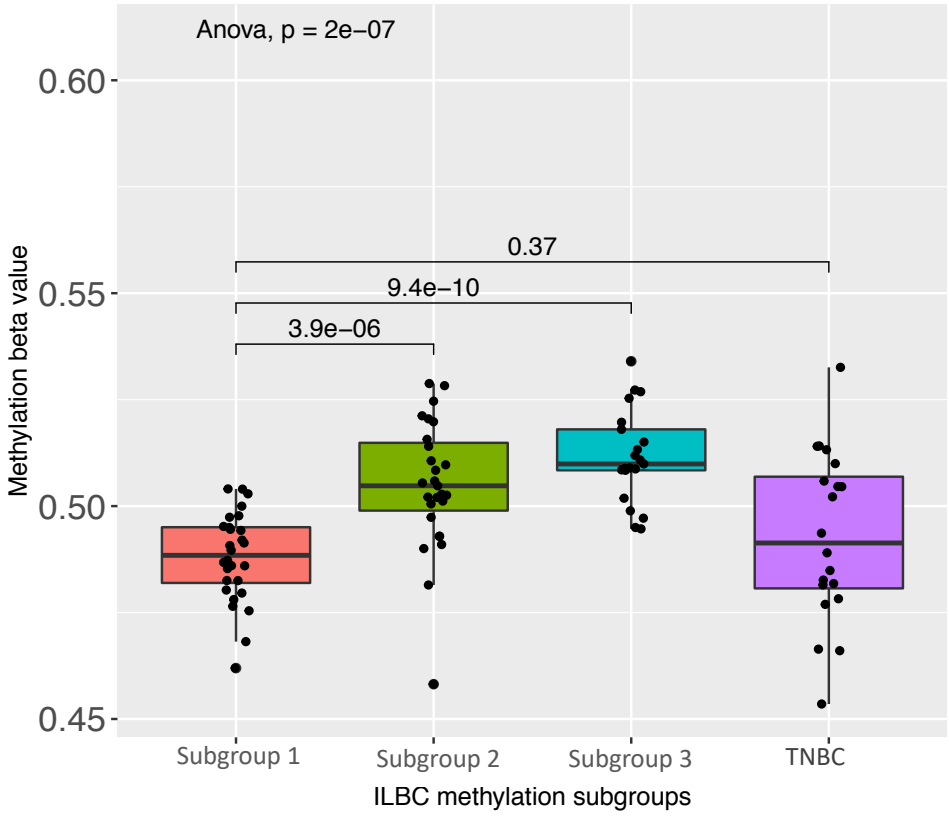


Figure 4.4: Average methylation level (beta-value) of ILBC methylation-defined subgroups and the triple negative breast cancer cases alongside Subgroup 1.

Box plots illustrating the average methylation level (beta-value), shown on the y-axis at all CpG positions genome-wide, of the ILBC methylation-defined subgroups; Subgroup 1, Subgroup 2 and Subgroup 3 and the triple negative breast cancer (TNBC) cases that were found to cluster alongside ILBC on the cluster dendrogram (shown on the x-axis). ANOVA *P*-values indicating the significance of the difference in average methylation levels for different comparisons are marked on the plots.

In terms of the genomic locations, the DMPs between Subgroup 1 and Subgroup 2 and between Subgroup 1 and Subgroup 3 were found to be largely distributed in the CpG island and the open sea region (region more than 4 kb away from the CpG island), whereas a smaller proportion was associated with the shore (up to 2 kb away from the CpG island) and the shelf (2-4 kb away from the CpG island) regions of the genome (Figure 4.5). The DMPs between Subgroup 2 and Subgroup 3 were predominantly associated with the CpG island region (85%) and a small proportion of DMPs were located in the shore (12%) and open sea region (3%) of the genome (Figure 4.5).

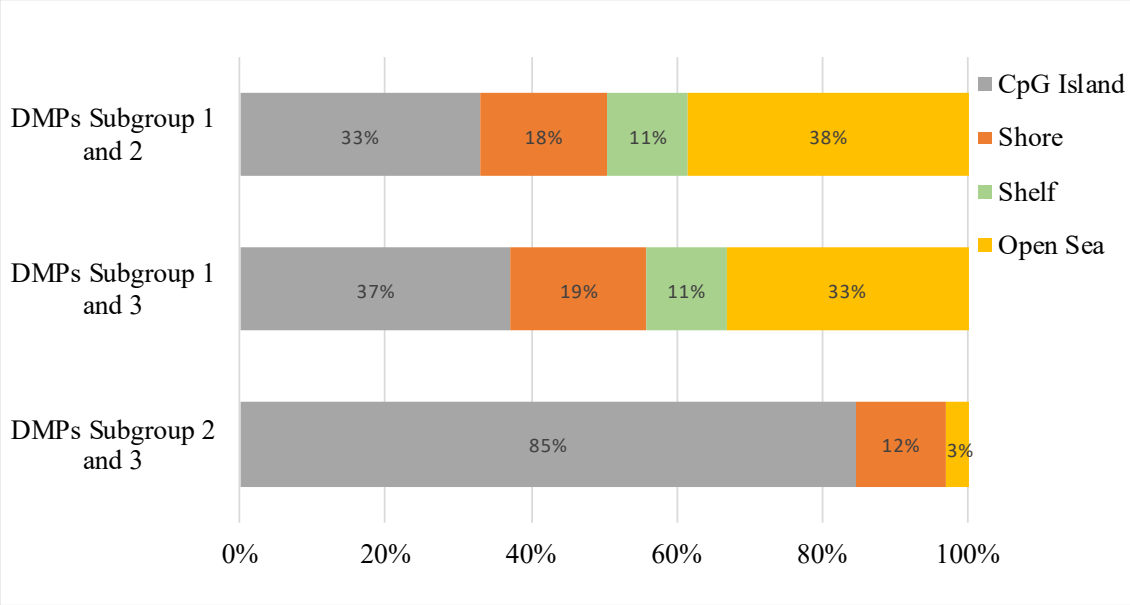


Figure 4.5: Genomic distribution of the differentially methylated positions.

Bar plot showing the distribution of the differentially methylated positions (DMPs) between the ILBC methylation-defined subgroups (shown on the y-axis) across different genomic regions; CpG island (region commonly associated with gene promoters), Shore (region up to 2 kb from the CpG island), Shelf (region up to 2-4 kb from the CpG island) and Open Sea (region more than 4 kb away from the CpG island). The x-axis shows the proportion of the DMPs in different genomic regions, represented by different bar colours as indicated in the legend on the top right corner.

The consecutive (located within 1000 bp of each other) methylated CpG positions or differentially methylated regions (DMRs) were identified between the ILBC methylation-defined subgroups as described in section 2.9.1 of the thesis. The DMPs between Subgroup 1 and Subgroup 2 clustered into 3,740 DMRs, of which 1,441 (39%) were overlapping with gene promoters and 2,299 (61%) were present in the intergenic regions. Between Subgroup 1 and Subgroup 3, 1,276 DMRs were identified, of which 498 (39%) were overlapping with gene promoters and 778 (61%) were present in the intergenic region. No DMRs were identified between Subgroup 2 and Subgroup 3 possibly due to the smaller number of DMPs between these two subgroups.

Table 4.2: Differentially methylated positions and differentially methylated regions identified in between the ILBC methylation-defined subgroups.

ILBC methylation-defined subgroup	Number of DMP	Number of DMR
Subgroup 1 and 2	27,675	3,740
Subgroup 1 and 3	13,067	1,276
Subgroup 2 and 3	65	0

ILBC: Invasive lobular breast cancer. DMP: Differentially methylated CpG position. DMR: Differentially methylated region.

The DMRs identified in Subgroup 1 were overlapping the genes that were mainly transcription factors and genes related to protein kinases activity. Among the most significant DMRs (*P*-value < 0.01), were the microRNAs such as *MIR433, MIR127, MIR136, MIR431* and *MIR432* and also included the genes related to immune response such as *TAP1*, *SOCS1*, *CD81*, *PSMB9,* and *TRIM27*. Gene ontology analysis performed on the top 10% (by *P*-value) of the DMRs between the subgroups showed that the genes overlapping the DMRs between Subgroup 1 and Subgroup 2 were significantly enriched (FDR-adjusted $P < 0.05$) for functional categories such as *Regulation of cytokine*

production involved in inflammatory response (GO:1900015), *Positive regulation of inflammatory response* (GO:0050729), *Regulation of leukocyte migration* (GO:0002685), *Cytokine secretion* (GO:0050663) and *Regulation of CD4-positive, alpha-beta T cell activation* (GO:2000514). Among the significantly enriched (FDR-adjusted $P < 0.05$) functional categories in the genes overlapping the DMRs between Subgroup 1 and Subgroup 3 were *Metabolism of RNA* (R-HSA-8953854), *mRNA Splicing* (R-HSA-72172), *Processing of Capped Intron-Containing Pre-mRNA* (R-HSA-72203), *RNA splicing via transesterification reactions* (GO:0000375) and *Ribonucleoprotein complex biogenesis* (GO:0022613). The ten most significant DMRs (by *P*-value) identified between the subgroups are summarised in Table 4.3 (between Subgroup 1 and Subgroup 2) and Table 4.4 (between Subgroup 1 and Subgroup 3). Table 4.5 summarises the ten most significant DMPs (by *P*-value) identified between Subgroup 2 and Subgroup 3.

Table 4.3: The ten most significant (by *P*-value) differentially methylated regions between ILBC methylation-defined Subgroup 1 and Subgroup 2.

Genomic location of the DMR[a] (hg19)	*P*-value*	Number of CpGs	Associated gene[†]	Protein Function	Reported relevance to cancer progression/etiology	Reference
chr17:1956958-1958162	$1.4x10^{-67}$	8	*HIC1*	Positive regulation of DNA damage response, signal transduction by p53 class mediator	Tumour suppressor function in several cancers including breast cancer.	(Cheng *et al.*, 2014; Y Wang *et al.*, 2018)
chr7:66368242-66369611	$8.6x10^{-58}$	7	*RP11-458F8.4*	Long non-coding RNA	-	-
chr6:32036449-32038027	$3x10^{-57}$	7	*NA*	-	-	-
chr6:32184410-32185984	$6.4x10^{-57}$	7	*NA*	-	-	-
chr14:67999752-67999935	$1.3x10^{-56}$	7	*PLEKHH1, TMEM229B*	-	-	-
chr6:32820546-32822017	$4x10^{-56}$	7	*TAP1, PSMB9*	*TAP1*-Antigen processing and presentation of exogenous peptide antigen via MHC class I, TAP-dependent. *PSMB9*-immunoproteasome	Downregulation of *TAP1* promotes immune evasion and associated with poor outcome.	(Pedersen *et al.*, 2017)
chr14:101348158-101351026	$5x10^{-56}$	7	*MIR433, MIR127, MIR136,*	Posttranscriptional gene silencing.	*MIR433*, *MIR127*, *MIR136*, *MIR431*- tumour suppressive role in breast cancer.	(S Wang *et al.*, 2014; M Yan *et al.*, 2016; X Hu *et*

			MIR431, RTL1, MIR432			*al.*, 2018; W Wang *et al.*, 2020)
chr16:11349372-11351330	$6.1x10^{-55}$	7	*SOCS1*	Negative regulation of cytokine signalling.	Oncogenic role in TNBC[b].	(Qian *et al.*, 2018)
chr6:31853565-31855709	$6.3x10^{-55}$	7	*EHMT2*	Histone H3-K27 methylation, negative regulation of G0 to G1 transition.	Promotes metastasis in breast cancer.	(K Kim *et al.*, 2018)
chr7:75701033-75703958	$1.2x10^{-54}$	7	*AC005077.14-001*	Pseudogene	-	-

[a] Differentially methylated region. * *P*-value computed using Stouffer's method (Stouffer *et al.*, 1949). [†] RefSeq gene name. NA: DMR not annotated to any gene. [b] Triple negative breast cancer.

Table 4.4: The ten most significant (by *P*-value) differentially methylated regions between ILBC methylation-defined Subgroup 1 and Subgroup 3.

Genomic location of the DMR[a] (hg19)	*P*-value*	Number of CpGs	Associated gene[†]	Protein Function	Reported relevance to cancer progression/ etiology	Reference
chr11:2397773-2398533	$1.0x10^{-56}$	7	*CD81*	Class I MHC mediated antigen processing and presentation, positive regulation of T-helper 2 cell cytokine production.	Increased expression is associated with poor patient prognosis in breast cancer.	(N Zhang *et al.*, 2018)
chr6:31597034-31599537	$9.2x10^{-54}$	7	*PRRC2A*	Regulation of pre-mRNA splicing.	-	-
chr16:88940941-88943452	$4.1x10^{-48}$	6	*NA*	-	-	-
chr6:28890887-28892849	$7.9x10^{-48}$	6	*TRIM27*	Negative regulation of interleukin-2 secretion	Promotes breast tumour growth	(Xing *et al.*, 2020)
chr7:66368242-66369525	$4.3x10^{-47}$	6	*RP11-458F8.4*	Long non-coding RNA	-	-
chr11:67260775-67262993	$2.2x10^{-42}$	5	*PITPNM1*	Phosphatidylinositol biosynthetic process.	Promotes epithelial to mesenchymal transition.	(Keinan *et al.*, 2014)
chr14:67999752-67999927	$1.1x10^{-41}$	5	*PLEKHH1, TMEM229B*	-	-	-
chr7:35734160-35734978	$1.1x10^{-41}$	5	*HERPUD2, RP11-379H18.1*	Cellular response to unfolded protein.	-	-
chr6:33422190-33422529	$1.4x10^{-41}$	5	*ZBTB9*	Transcription regulation.	ER-alpha target gene.	(Lin *et al.*,

						2007)
chr6:32820577-32822017	$7.4x10^{-40}$	5	*TAP1, PSMB9*	*TAP1*-Antigen processing and presentation of exogenous peptide antigen via MHC class I, TAP-dependent. *PSMB9*- immunoproteasome	Downregulation of *TAP1* promotes immune evasion and associated with poor outcome.	(Pedersen *et al.*, 2017)

[a] Differentially methylated region. **P*-value computed using Stouffer's method (Stouffer *et al.*, 1949). [†] RefSeq gene name. NA: DMR not annotated to any gene.

Table 4.5: The ten most significant (by *P*-value) differentially methylated positions between Subgroup 2 and Subgroup 3.

DMP[a]	*P*-value*	Associated gene[†]	Protein Function	Reported relevance to cancer progression/ etiology	Reference
cg11715828	2.4x10-6	*RNF220*	Protein ubiquitination, positive regulation of canonical Wnt signalling pathway.	Role in Wnt-related tumourigenesis such as colon cancer.	(P Ma *et al.*, 2014)
cg01656394	2.2x10-5	*NA*	-	-	-
cg15289658	2.2x10-5	*PSAT1*	Pyridoxal phosphate binding	Associated with response to endocrine therapy in breast cancer.	(Martens *et al.*, 2005)
cg12796383	4.9x10-5	*NA*	-	-	-
cg18529845	5.8x10-5	*SRD5A2*	Metabolism of steroid hormones	Regulates progesterone metabolism in breast cancer.	(Lewis *et al.*, 2004)
cg16896847	6.0x10-5	*MAFA*	DNA binding and transcription factor activity	-	-
cg22674412	1.1x10-4	*RPRML*	p53 dependent arrest of the cell cycle.	Tumour suppressor activity, inhibits cell migration and invasion.	(Buchegger *et al.*, 2016)
cg04974587	1.1x10-4	*ARHGAP20*	GTPase activator activity.	-	-
cg14305313	1.2x10-4	*NA*	-	-	-
cg24879335	1.3x10-4	*TF*	Iron binding protein	-	-

[a] Differentially methylated position. * *P*-value calculated using moderated t.test. [†] RefSeq gene name. NA: DMP not annotated to any gene.

4.3.3 Correlation with clinicopathological features

To further investigate any possible association between the ILBC methylation-defined subgroups and the clinicopathological features of the tumour samples, the subgroups were evaluated using available clinical and pathological data and epidemiological information (including family history). A detailed comparison of the ILBC methylation-defined subgroups based on tumour characteristics and family history data are shown in Table 4.6. The associations were tested using Pearson's chi-square test and ANOVA for categorical and continuous variables, respectively.

Table 4.6: Clinical and pathological features of the ILBC methylation-defined subgroups.

ILBC features	Subgroup 1 (n = 28)	Subgroup 2 (n = 27)	Subgroup 3 (n = 21)	*P*-value*
Median age at diagnosis, years [interquartile range]	67 [25%; 60]	62 [25%; 59]	60 [25%; 51]	0.36
Age group, n (%)				
<50	3 (11)	1 (4)	2 (10)	0.23
50-60	4 (14)	8 (30)	9 (43)	
60+	20 (71)	17 (63)	10 (48)	
Missing	1 (4)	1 (4)	0 (0)	
Study, n (%)				
MCCS[a]	18 (64)	24 (89)	17 (81)	0.28
kConFab[b]	3 (11)	1 (4)	1 (5)	
ABCFR[c]	7 (25)	2 (7)	3 (14)	
Ethnicity, n (%)				
Australian/NZ[d]	18 (64)	22 (82)	13 (62)	0.41
English	0 (0)	1 (4)	1 (5)	
Greek/Italian/Maltese	3 (11)	1 (4)	3 (14)	
Missing	7 (25)	3 (11)	4 (19)	
Year of cancer diagnosis, n (%)				
1992-1997	6 (21)	8 (30)	6 (29)	0.52
1998-2003	9 (32)	8 (30)	9 (43)	
2004-2008	6 (21)	8 (30)	5 (24)	
2009 and later	7 (25)	3 (11)	1 (5)	
Tumour ER expression, n (%)				
Positive	27 (96)	23 (85)	17 (81)	0.19
Negative	0 (0)	3 (11)	1 (5)	
Missing	1 (4)	1 (4)	3 (14)	
Tumour PR expression, n (%)				
Positive	18 (64)	19 (70)	13 (62)	0.95
Negative	8 (29)	7 (26)	5 (24)	
Missing	2 (7)	1 (4)	3 (14)	
Tumour HER2 expression, n (%)				
Positive	2 (7)	1 (4)	4 (19)	0.25
Negative	19 (68)	16 (59)	13 (62)	
Missing	7 (25)	10 (37)	4 (19)	
ER positive/PR positive, n (%)				
Yes	18 (64)	19 (70)	13 (62)	0.95
No	10 (36)	8 (30)	8 (38)	
ER positive/PR negative, n (%)				
Yes	8 (29)	4 (15)	4 (19)	0.41
No	20 (71%)	23 (85)	17 (81)	

ER negative/PR negative, n (%)				
Yes	0 (0)	3 (11)	1 (5)	0.19
No	28 (100)	24 (89)	20 (95)	
ICDO[†]-code, n (%)				
8520	20 (71)	24 (89)	14 (67)	0.16
8522	7 (25)	3 (11)	7 (33)	
Missing	1 (4)	0 (0)	0 (0)	
Tumour Grade, n (%)				
Grade I	3 (11)	1 (4)	2 (10)	0.71
Grade II	16 (57)	18 (67)	13 (62)	
Grade III	4 (14)	3 (11)	1 (5)	
Missing	5 (18)	5 (19)	5 (24)	
Tumour Stage, n (%)				
1A/1B	9 (32)	15 (56)	10 (48)	0.39
2A/2B	4 (14)	6 (22)	6 (29)	
3A/3C	5 (18)	2 (7)	1 (5)	
4	0 (0)	1 (4)	0 (0)	
Missing	10 (36)	3 (11)	4 (19)	
Median tumour Purity, % [interquartile range]	56 [25%, 50]	54 [25%, 52]	58 [25%, 55]	0.73
Tumour size (mm), n (%)				
<5	1 (4)	0 (0)	0 (0)	0.52
5-10	3 (11)	4 (15)	4 (19)	
10-20	10 (36)	13 (48)	8 (38)	
20-30	2 (7)	2 (7)	4 (19)	
35-40	3 (11)	3 (11)	0 (0)	
40+	2 (7)	3 (11)	0 (0)	
Missing	7 (25)	2 (7)	5 (24)	
Laterality, n (%)				
Left	13 (46)	9 (33)	7 (33)	0.51
Right	14 (50)	17 (63)	14 (67)	
Bilateral	0 (0)	1 (4)	0 (0)	
Missing	1 (4)	0 (0)	0 (0)	
Multifocal tumours, n (%)				
Yes	4 (14)	11 (41)	5 (24)	0.11
Missing	24 (86)	16 (59)	16 (76)	
Recurrence status, n (%)				
Yes	3 (11)	1 (4)	1 (5)	0.33
No	13 (46)	20 (74)	14 (67)	
Missing	12 (43)	6 (22)	6 (0)	
Age at menarche, n (%)				
<12	5 (18)	6 (22)	2 (10)	0.45
12-14	10 (36)	17 (63)	12 (57)	
14+	3 (11)	1 (4)	3 (14)	
Missing	0 (0)	3 (11)	4 (19)	

Parity and lactation history, n (%)				
Nulliparous	2 (7)	1 (4)	1 (5)	0.50
Parous, lactated	19 (68)	24 (89)	16 (76)	
Parous, never lactated	0 (0)	0 (0)	1 (5)	
Missing	7 (25)	2 (7)	3 (14)	
Hormone replacement Therapy, n (%)				
Yes	9 (32)	12 (44)	6 (29)	0.51
No	17 (61)	15 (56)	15 (71)	
Missing	2 (7)	0 (0)	0 (0)	
Oral contraceptive use, n (%)				
Yes	8 (29)	19 (70)	13 (62)	**0.02**
No	13 (46)	6 (22)	5 (24)	
Missing	7 (25)	2 (7)	3 (14)	
Father had cancer, n (%)				
Yes	6 (21)	7 (26)	4 (19)	0.81
No	12 (43)	17 (63)	13 (62)	
Missing	10 (36)	3 (11)	4 (19)	
Mother had cancer, n (%)				
Yes	3 (11)	3 (11)	8 (38)	**0.03**
No	15 (54)	21 (78)	9 (43)	
Missing	10 (36)	3 (11)	4 (19)	
Female relative with breast cancer, n (%)				
Yes	3 (11)	9 (33)	6 (29)	0.22
No	13 (46)	11 (41)	8 (38)	
Missing	12 (43)	7 (26)	7 (33)	

ILBC: Invasive lobular breast cancer. * *P*-values are for chi-square test and ANOVA test for categorical and continuous variables, respectively. [a] Melbourne Collaborative Cohort Study. [b] The Kathleen Cuningham Foundation Consortium for Research into Familial Breast Cancer. [c] Australian Breast Cancer Family Registry. [d] New Zealand. ER: Estrogen receptor. PR: Progesterone receptor. HER2: Human epidermal growth factor receptor 2. [†] International Classification of Diseases for Oncology (8520- Invasive lobular breast cancer, 8522- Infiltrating ductal and lobular carcinoma), ER, PR and HER2 expression status was determined using immunohistochemistry as described in (Blows *et al.*, 2010).

The median age of women at cancer diagnosis was 67 years for Subgroup 1, 62 years for Subgroup 2 and 60 years for Subgroup 3. The women were primarily of Australian/New Zealand ethnic background (64% in Subgroup 1, 82% in Subgroup 2 and 62% in Subgroup 3). As previously observed in the unsupervised cluster analysis (Figure 4.1), the clustering of ILBCs was not significantly driven by the ER, PR and HER2 receptor expression status of the tumour and the majority of samples in all three subgroups were ER and PR positive and HER2 negative (Figure 4.6). This was consistent with what is known for ILBC (Ciriello *et al.*, 2015; Reed *et al.*, 2015). There was no ER negative case in Subgroup 1, whereas in Subgroup 2, 3/27 (11%) samples were ER negative and in Subgroup 3, 1/21 (5%) sample was ER negative. Subgroup 1 was enriched for ER positive/PR negative tumours, 8/28 (29%) compared with Subgroup 2, 4/27 (15%) and Subgroup 3, 4/21 (19%) however it was not statistically significant. There were 2 triple-negative breast tumours, both in Subgroup 2. Out of the two triple-negative tumours in Subgroup 2, one was associated with DCIS and the woman was premenopausal. Both tumours were multifocal and displayed classic ILBC morphological subtype (ICDO-code-8520).

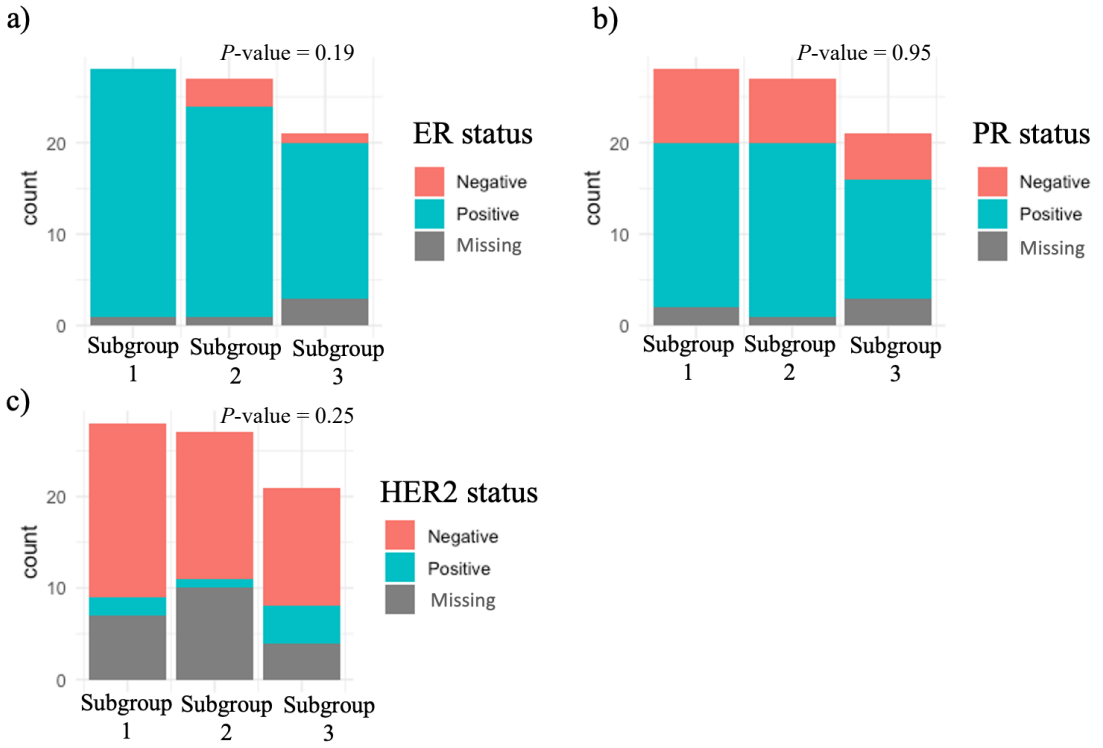


Figure 4.6: Distribution of hormone receptor and human epidermal growth factor receptor 2 expression status across the ILBC methylation-defined subgroups.

The bar plots showing the distribution of hormone receptor expression status measured using immunohistochemistry **a**) Estrogen receptor (ER); **b**) Progesterone receptor (PR) and **c**) Human epidermal growth factor receptor 2 (HER2) in the ILBC methylation-defined subgroups (shown on the x-axis). The sample count is shown on the y-axis. Respective colours representing the hormone receptor expression status are marked in the legend besides each bar plot.

Grade II was the most frequently observed tumour grade across the three subgroups: 16/28 (57%) of cases in Subgroup 1, 18/27 (67%) of cases in Subgroup 2 and 13/21 (62%) of cases in Subgroup 3 (Figure 4.7a). Four cases of grade III tumour were recorded in Subgroup 1 compared with three cases in Subgroup 2 and one case in Subgroup 3. Tumour stage was recorded as stage 1A/1B, 2A/2B, 3A/3C and stage 4 with stage 1A/1B being the most frequently observed across the subgroups (Figure 4.7b). In Subgroup 2, 15/27 (56%) of cases and in Subgroup 3, 10/21 (48%) of cases were stage 1A/1B tumour compared with 9/28 (32%) of cases in Subgroup 1. Advanced stage tumours (3A/3C and stage 4) were relatively rare with five cases of tumour stage 3A/3C in Subgroup 1 (2/5 cases showed mixed lobular/ductal morphology, ICDO-code 8522), two cases each in Subgroup 2 and one case in Subgroup 3 (all showing classic ILBC morphology, ICDO-code-8520). There was one case of stage 4 tumour that clustered in Subgroup 2 (ICDO-code-8520).

The size of the tumour at diagnosis ranged from 5 mm to more than 40 mm across the ILBC subgroups with most cases showing tumour sizes ranging from 10 mm to 20 mm (10/28, 36% in Subgroup 1; 13/27, 48% in Subgroup 2 and 8/21, 38% in Subgroup 3) (Figure 4.7c). Five cases had large size tumours (>40 mm), out of which, two cases were from Subgroup 1 and three cases were from Subgroup 2.

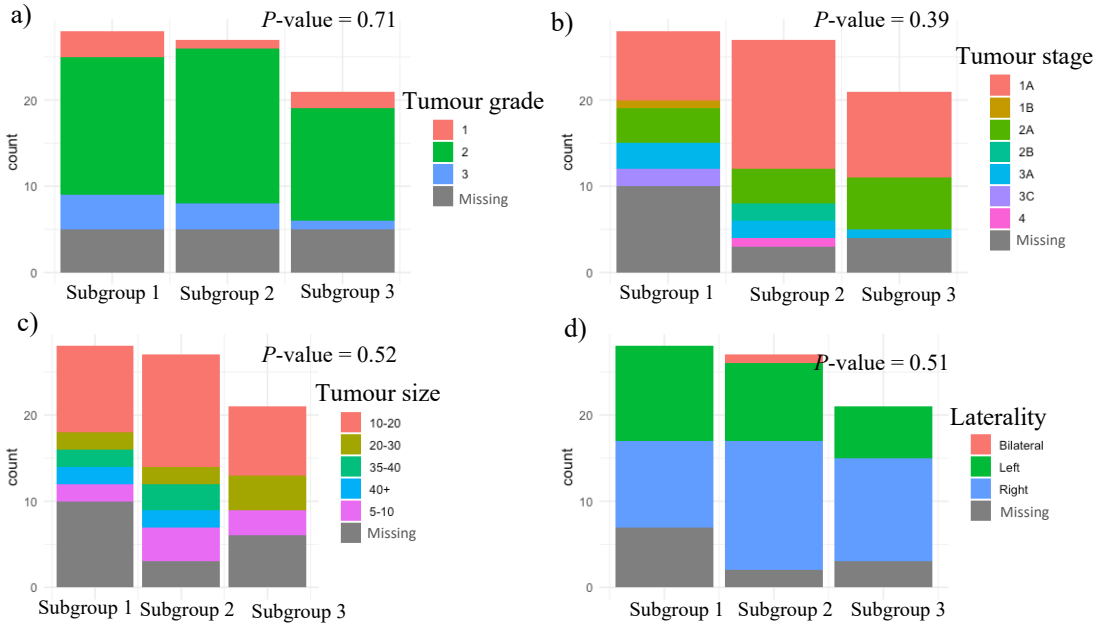


Figure 4.7: Distribution of tumour features across the ILBC methylation-defined subgroups.

Bar plots showing the distribution of tumour features **a**) Tumour grade; **b**) Tumour stage; **c**) Tumour size and **d**) Tumour laterality in ILBC methylation-defined subgroups (shown on the x-axis). The sample count is shown on the y-axis. Respective colours depicting different tumour features are shown in the legend besides each bar plot.

ILBCs are known to be more frequently bilateral and multifocal compared with other breast cancer types (Arpino *et al.*, 2004). Across the ILBC methylation-defined subgroups, Subgroup 1 had equal proportion of women with tumour in left and right breasts, whereas Subgroup 2 and Subgroup 3 had higher proportion of women with tumours in the right breast compared to the left breast (Figure 4.7d). There was one bilateral case that clustered in Subgroup 2. There was an enrichment of multifocal tumours in Subgroup 2 (11/27, 41%) compared with Subgroup 1 (4/28, 14%) and Subgroup 3 (5/21, 24%) (Figure 4.8). No significant association was observed between ILBC methylation subgroups and any of the tumour features discussed above (tumour grade, *P*-value = 0.71; tumour stage, *P*-value = 0.39; tumour size, *P*-value = 0.52; tumour laterality, *P*-value = 0.51 and multifocality, *P*-value = 0.11).

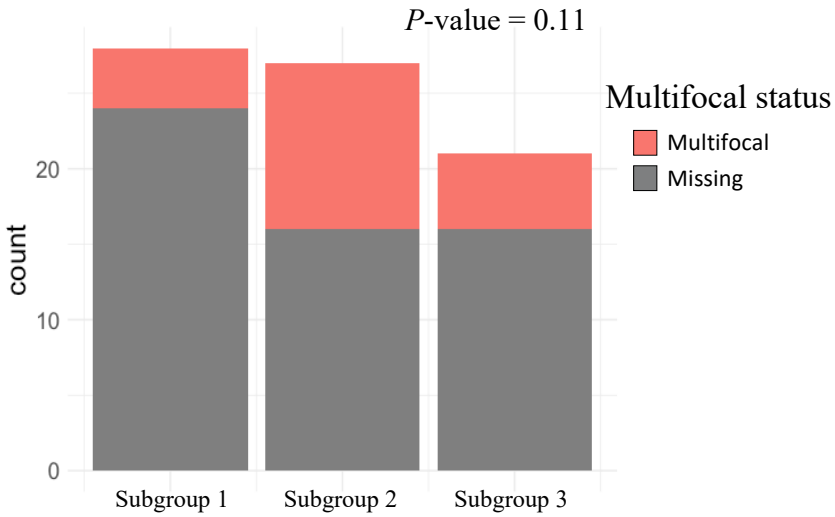


Figure 4.8: Distribution of tumour focality status across the ILBC methylation-defined subgroups.

The bar plot shows the distribution of tumour focality status in the ILBC methylation-defined subgroups (shown on the x-axis) and the sample count shown on the y-axis. Respective colours depicting the multifocal status are shown in the legend besides the bar plot.

Factors such as use of oral contraceptive, hormone replacement therapy, age at menarche and parity status are well known risk factors for breast cancer (Clavel-Chapelon & Gerber, 2002). In terms of oral contraceptive pill usage, Subgroup 2 and Subgroup 3 showed a significant enrichment (*P*-value = 0.02) for women who had a history of oral contraceptive pills usage (19/27, 70% in Subgroup 2 and 13/21, 62% in Subgroup 3) compared with Subgroup 1 where 8/28 (29%) of the women had used or had been using oral contraceptive pills (Figure 4.9a). In terms of hormone replacement therapy (HRT), Subgroup 2 was again found to be slightly enriched for women who have taken HRT (12/27, 44%) compared with Subgroup 1 (9/28, 32%) and Subgroup 3 (6/21, 29%), however, this association was not statistically significant (*P*-value = 0.51) (Figure 4.9b).

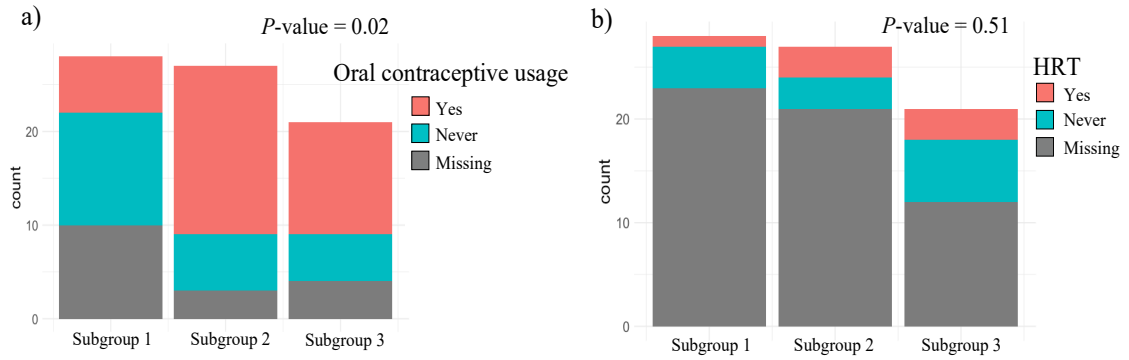


Figure 4.9: Distribution of women reproductive history and other breast cancer risk factors across the ILBC methylation-defined subgroups.

Bar plots showing the distribution of reproductive history and other breast cancer risk factors **a**) Oral contraceptive usage and **b**) Hormone replacement therapy (HRT) across the ILBC methylation-defined subgroups (shown on the x-axis). The sample count is shown on the y-axis. Respective colours depicting the different factors are shown in the legend besides each bar plot.

A complex crosstalk exists between genetic variations and epigenetic patterns in tumours (You & Jones, 2012). Genetic variations are known to alter the DNA methylation patterns (Kerkel *et al.*, 2008; Bell *et al.*, 2011). Therefore, we were interested to know if any inherited genetic factor was driving the ILBC subgroups and to find out more, the subgroups were evaluated for the family history data of the women in the three ILBC methylation subgroups. Subgroup 3 was found to have a significant enrichment (*P*-value = 0.03) for women who had a mother with history of cancer (any kind) (Figure 4.10a). Looking at the data specific for breast cancer, Subgroup 2 and Subgroup 3 showed an enrichment for women who had a female relative with breast cancer (9/27, 33% of women in Subgroup 2 and 6/21, 28% of women in Subgroup 3) compared with Subgroup 1 (3/28, 11%) however, the difference was not statistically significant (*P*-value = 0.22) (Figure 4.10c). No significant association was found between the subgroups and women who had a father with history of cancer (any kind) (Figure 4.10b).

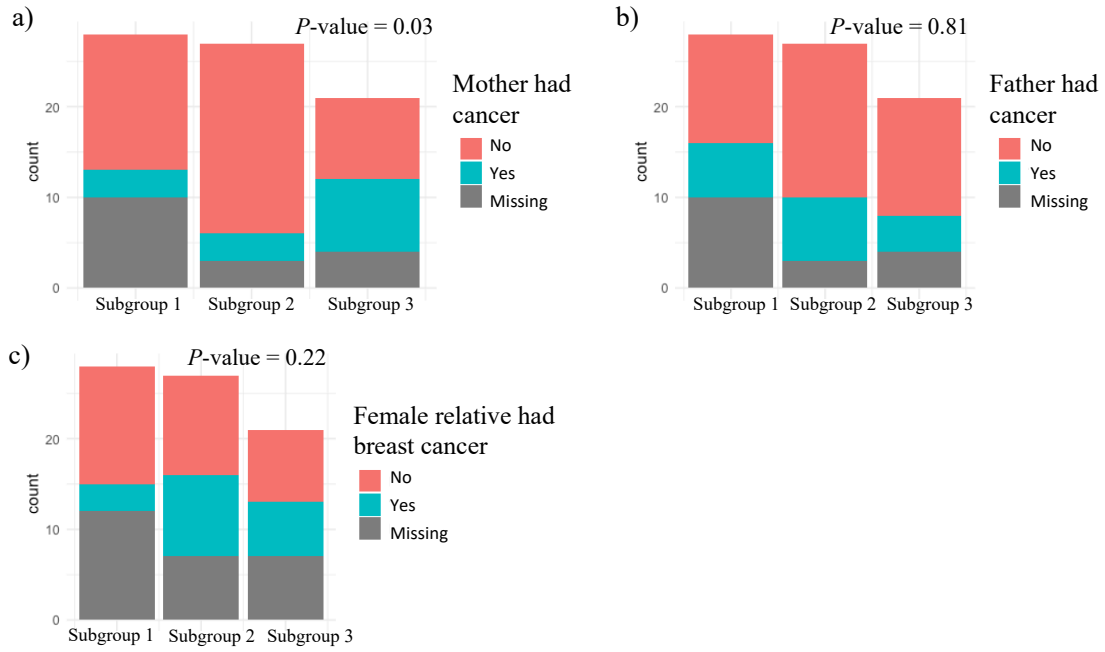


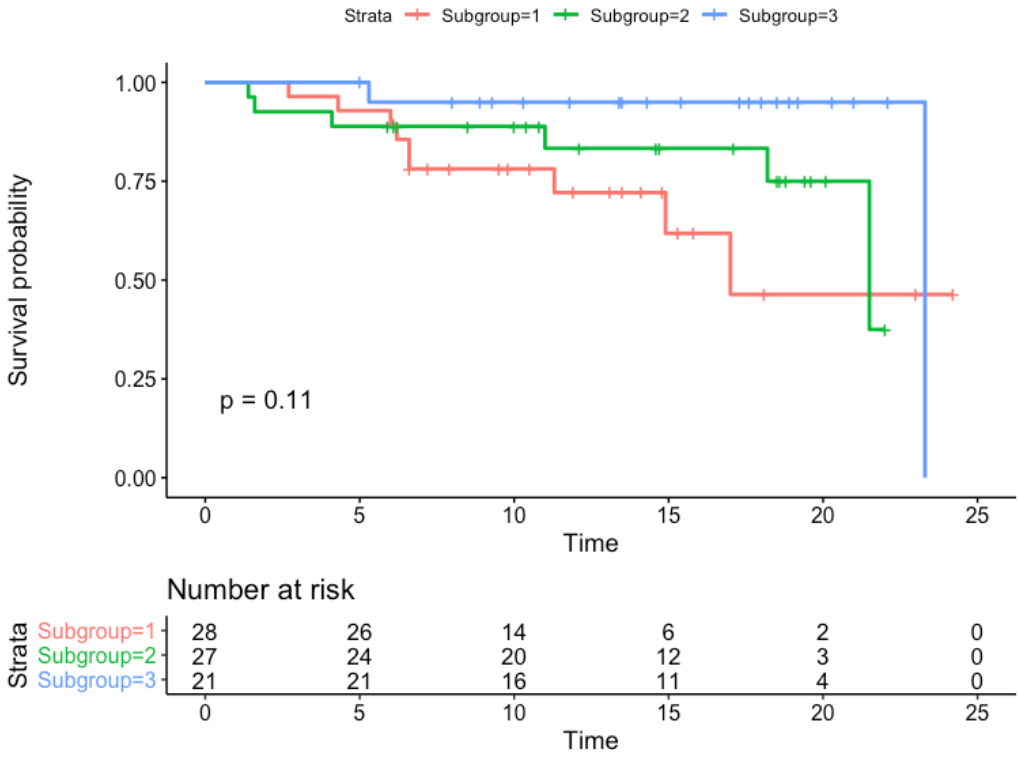
Figure 4.10: Distribution of family history information of women across the ILBC methylation-defined subgroups.

The bar plots show the distribution of the family history information of women with ILBC **a**) history of mother with cancer; **b**) history of father with cancer and **c**) history of a female relative with breast cancer across the ILBC methylation-defined subgroups (shown on the x-axis. The sample count is shown on the y-axis. Respective colours depicting the different factors are shown in the legend besides each bar plot.

4.3.4 Methylation subgroups differ in their overall survival

As information on breast cancer specific survival was not available, difference in overall survival was tested among the ILBC methylation-defined subgroups. During the median follow-up period of 13 years, there were 17 confirmed deaths recorded in the sample set. Kaplan-Meir survival curves stratified according to the ILBC methylation-defined subgroups are shown in the Figure 4.11a. The survival probability of the three subgroups were compared using a log-rank test that showed that Subgroup 1, with a relatively hypomethylated profile among the three subgroups, showed a poor overall survival in comparison with Subgroup 2 and Subgroup 3 (Figure 4.11a). While comparing the overall survival probability between the two most distinct methylation-defined subgroups, i.e., Subgroup 1 (the most hypomethylated) and Subgroup 3 (the most hypermethylated), Subgroup 1 showed a significantly reduced overall survival compared with Subgroup 3 (log-rank test, *P*-value = 0.035) (Figure 4.11b). After adjusting for age and year of diagnosis, in a multivariable analysis using Cox proportional regression hazard model, the association remained consistent, with Subgroup 1 showing a worse overall survival compared with Subgroup 2 and Subgroup 3 (Table 4.7).

a)


Strata Subgroup=1 Subgroup=2 Subgroup=3
Survival probability
1.00
0.75
0.50
0.25
0.00
p = 0.11
0 5 10 15 20 25
Time
Number at risk
Strata
Subgroup=1 28 26 14 6 2 0
Subgroup=2 27 24 20 12 3 0
Subgroup=3 21 21 16 11 4 0
0 5 10 15 20 25
Time

b)

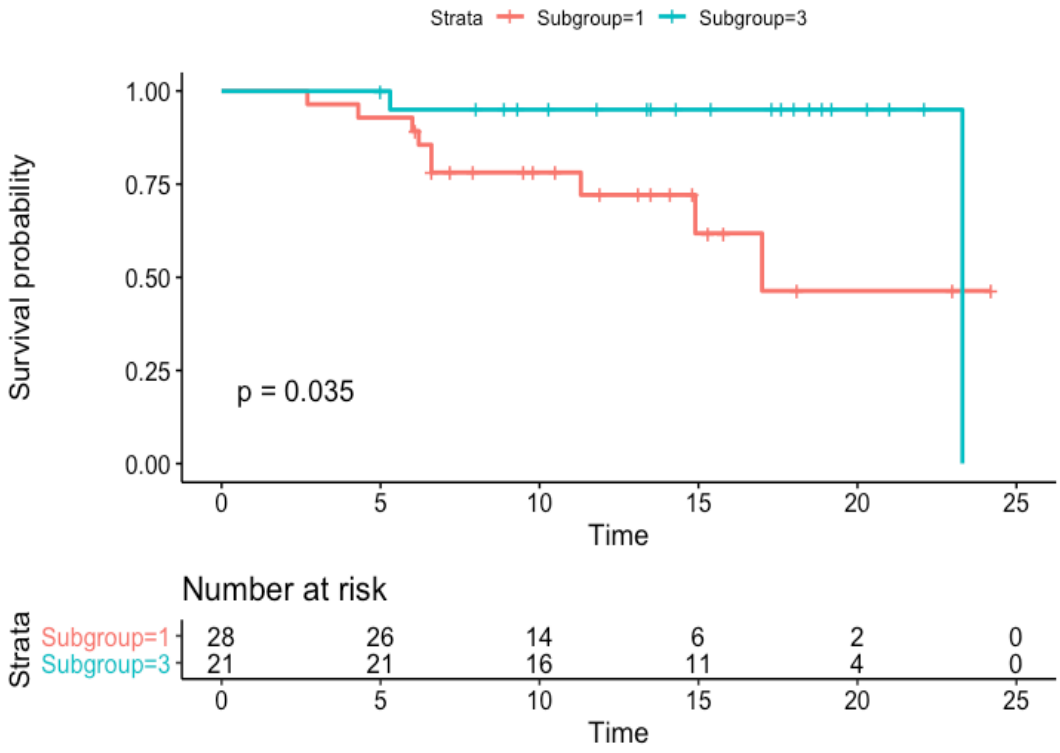

Strata Subgroup=1 Subgroup=3
Survival probability
1.00
0.75
0.50
0.25
0.00
p = 0.035
0 5 10 15 20 25
Time
Number at risk
Strata
Subgroup=1 28 26 14 6 2 0
Subgroup=3 21 21 16 11 4 0
0 5 10 15 20 25
Time

Figure 4.11: Kaplan-Meier plot stratified according to the ILBC methylation-defined subgroups.

Kaplan-Meier plots show the overall survival curves of **a**) the ILBC methylation-defined subgroups; Subgroup 1, Subgroup 2 and Subgroup 3 and **b**) the ILBC methylation-defined subgroups; Subgroup 1 and Subgroup 3. Total follow-up time (number of years) is shown on the x-axis and the survival probability (the probability of patients surviving past a specific time) is shown on the y-axis. The survival curves of the three ILBC methylation subgroups; Subgroup 1, Subgroup 2 and Subgroup 3 are represented by red, green and blue lines, respectively. Vertical lines on the survival curves indicate censored data and their corresponding x-values indicate the time at which the censoring occurred. Table below the plot shows samples at risk over time.

Table 4.7: Hazard ratios for the association between the invasive lobular breast cancer methylation-defined subgroups and overall survival.

Reference Subgroup*	**Comparison subgroup**	**Unadjusted**		**Adjusted for age and year of diagnosis**	
		HR (95%CI)	***P*-value**	**HR (95%CI)**	***P*-value**
Subgroup 1	**Subgroup 2**	0.64 (0.22 -1.86)	0.41	0.59 (0.19-1.79)	0.35
	Subgroup 3	0.22 (0.05-1.05)	0.05	0.16 (0.03-0.88)	0.03

* The subgroup that served as a reference to calculate the hazard ratios. HR: Hazard ratio. CI: Confidence interval.

4.4 Discussion

This chapter provides data to support the hypothesis that subgroups of ILBC with increased DNA methylation homogeneity can be identified using genome-wide tumour DNA methylation measurement. This approach defined three subgroups of ILBC via unsupervised cluster analysis that had significantly differentially methylated positions, some important differences in epidemiological risk factors and provided some evidence for differences in patient prognosis.

4.4.1 ILBC subgroups associated with a family history of breast cancer

A strong heritable component for ILBC susceptibility has been reported in previous studies. Allen-Brady et al., (2005) reported that relatives of women affected by ILBC had an increased risk for ILBC (first-degree relative: familial relative risk (FRR) = 4.51, 95% CI: 2.8-6.9) and an increased risk for any type of breast cancer (first-degree relative: FRR = 2.5, 95% CI: 2.1-2.9) (Allen-Brady *et al.*, 2005). This was supported by a recent report from Henry & Cannon-Albright, (2019) who identified significant familial clustering in ILBC with an increased RR for breast cancer of any histology in second-degree relatives of women with ILBC estimated to be 1.36 (95% CI: 1.25-1.47), *P*-value <0.001 (Henry & Cannon-Albright, 2019).

In this study, Subgroup 3 showed a significant enrichment (*P*-value = 0.03) for women who had a mother with history of cancer (any kind). Subgroups 2 and 3 were also found to be enriched for women who had a female relative with breast cancer. Genetic variation is known to influence DNA methylation both at the site of genetic variation and in some instances considerable distance from the genetic variation (Kerkel *et al.*, 2008; Bell *et al.*, 2011). The mechanisms behind this observation include simple direct

disruption (or creation) of the CpG sites, genetic disruption of the methylation machinery through to disruption of complex regulatory interactions where the effect is considerably distant from the genetic variation (McRae *et al.*, 2014; Taylor *et al.*, 2019). Data from this study could suggest that women in Subgroups 2 and 3 carry germline genetic variation and/or could have acquired common somatic genetic variation that influences the methylation pattern of their ILBC. The germline genetic variation could potentially be a rare variant that influences the tumourigenic pathway sufficiently to generate the methylation patterns described in the subgroups or could be a reflection of heritable common genetic variation that influences the methylation status of the ILBCs. Samples from Subgroup 2 and Subgroup 3 should be investigated further for possible association with a shared heritable component. This could possibly lead to the identification of subgroup-specific novel predisposition genes and or common genetic variation associated with ILBC.

4.4.2 Genome-wide DNA methylation of mixed lobular ductal histological subtype

ILBC can be morphologically challenging to categorise as some ILBC tumours also have growth patterns typical of ductal carcinomas. These tumours are classified as mixed lobular ductal carcinoma and are defined as having a lobular pattern in at least 50% of the tumour and a ductal pattern in 10-50% of the tumour (Lakhani, 2012). Mixed lobular ductal subtype (ICDO-code, 8522) constituted 17/76 (22%) of the ILBC tumours in the three methylation subgroups. The ILBC cases with mixed lobular ductal morphological features were found to cluster in the three subgroups: 7/28, 25% of cases in Subgroup 1, 3/27, 11% of cases in Subgroup 2 and 7/21, 33% of cases in Subgroup 3. A higher proportion of grade III tumours were observed in the mixed lobular ductal

subtype compared with the classic ILBC subtype (4/17, 24% grade III tumours in mixed lobular ductal *versus* 4/59, 7% in classic ILBC, chi-square, *P*-value = 0.07). Women who had a premenopausal cancer diagnosis showed a higher proportion of mixed lobular ductal tumours (3/4, 75% mixed lobular ductal tumour *versus* 1/4, 25% classic ILBC, chi-square, *P*-value = 0.06). Although mixed lobular ductal subtype has been associated with more aggressive tumour features (Rakha *et al.*, 2009; Arps *et al.*, 2013), no significant difference in age at diagnosis (t.test, *P*-value = 0.09) and other prognostic features such as stage (chi square, *P*-value = 0.87) and tumour size (t.test, *P*-value = 0.37) was observed between classic ILBC and the mixed lobular-ductal type in this study. Although, a separation of mixed lobular ductal tumours into hypermethylated and hypomethylated subgroups was expected, the data did not suggest such differences in their methylation profiles. Some evidence relating to this is presented in the recent TCGA study where two groups of mixed lobular ductal tumours are defined namely “ILC-like” and “IDC-like” resembling ILBC and IDBC, respectively at their genomic level (Ciriello *et al.*, 2015), which potentially could influence their genome-wide methylation profiles. Further study is needed to better characterise and further refine the classification of mixed lobular-ductal tumours that should include molecular pathology approaches that can consider the growth patterns independently (rather than in a single tumour-enriched DNA sample). Single cell approaches may be useful here.

4.4.3 Hypomethylation and Subgroup 1

Subgroup 1 displayed a distinct methylation profile compared with the other two ILBC methylation-defined subgroups. Samples in Subgroup 1 had a hypomethylated profile across the DMPs identified between the subgroups as well as across all CpG positions genome-wide (global methylation, i.e., average across the genome) Figure 4.3.

The clustering of Subgroup 1 in the dendrogram was found to be alongside the TNBC cases. The TNBC cases and Subgroup 1 belonged to two separate main branches in the dendrogram; Group A and Group B, respectively on Figure 4.1. The global methylation profile (average methylation beta-value genome-wide) of Subgroup 1 resembled the TNBC cases (ANOVA, *P*-value = 0.37) compared with Subgroup 2 (ANOVA, *P*-value = $3.9x10^{-6}$) and Subgroup 3 (ANOVA, *P*-value = $9.4x10^{-10}$) (Figure 4.4). This reveals a subset of ILBC cases that may share some features with TNBC. Although, none of the samples in Subgroup 1 were TNBC (based on their hormone receptor expression status), a higher proportion of ER positive PR negative tumours were found in Subgroup 1 compared to the other two subgroups (8/28, 29% in Subgroup 1 *versus* 4/27, 15% in Subgroup 2 and 4/21, 19% in Subgroup 3). Although the difference was not statistically significant (*P*-value = 0.41), it may be reflective of a small sample size. PR negative tumours are known to be associated with poor patient outcome (Rakha, Reis-Filho, & Ellis, 2010) which is consistent with our finding that also associated Subgroup 1 with poorer overall survival when compared with Subgroup 2 and Subgroup 3. The proportion of stage 3A/3C tumours were higher in Subgroup 1 (5/28, 18% *versus* 2/27, 7% in Subgroup 2 and 1/21, 5% in Subgroup 3, *P*-value = 0.19), but was not a reflection of the distribution of mixed ductal ILBC in these subgroups. We found 3/28 (11%) cases in Subgroup 1 showing recurrence (one case of distant and two cases of local recurrences) compared with 1/27 (4%) case of distant recurrence in Subgroup 2 and 1/21 (5%) case of distant recurrence in Subgroup 3. These features associated with aggressive tumour behaviour further suggesting a similarity in clinical behaviour and suggests that Subgroup 1 may represent an aggressive form of ILBC tumours independent of histology.

4.4.4 Implication of the immune response in Subgroup 1 tumourigenesis

The DMRs between the subgroups were mainly overlapping with genes that were transcription factors and genes related to protein kinase and protein transferase activity. The DMRs between Subgroup 1 and Subgroup 2 were mainly enriched in biological pathways related to immune system regulation such as *Regulation of cytokine production involved in inflammatory response* (GO:1900015), *Positive regulation of inflammatory response* (GO:0050729), *Regulation of leukocyte migration* (GO:0002685), *Cytokine secretion* (GO:0050663), *Regulation of CD4-positive, alpha-beta T cell activation* (GO:2000514) and involved genes such as *MAPK14*, *IL17RA*, *APPL1*, *GPSM3*, *ADAM8*, *CD81* and *HYAL2*. This involvement of immune response related pathways and genes differentially methylated between Subgroup 1 and Subgroup 2 might suggest a difference at the level of immune regulation that might in part differentiate these subgroups. This finding is consistent with the gene-expression based subtypes of ILBC identified in previous studies that performed cluster analysis using mRNA gene expression data of 144 and 106 ILBC samples, respectively (Ciriello *et al.*, 2015; Michaut *et al.*, 2016). Both these studies defined one of the ILBC subtypes as "immune related". Michaut et al., (2016) defined this subtype characterised by upregulation of genes associated with cytokine/chemokine signalling and showed a higher lymphocytic infiltration. Similarly, Ciriello et al., (2015) also defined a subtype of ILBC characterised by overexpression of interleukin and chemokine mRNA expression and also showed an overexpression of macrophage associated signalling. This suggests that there might be functional difference in the immune activity that define the ILBC subgroups and that this subset of ILBC may benefit from targeted therapies involving immune response and should be further investigated.

4.4.5 Limitations of the study

Despite some promising results, one of the possible limitations of this study is the hierarchical clustering approach used to define the methylation subgroups. Clustering algorithms work by linking the two most similar samples together first, and then successively merging other samples in the order of similarity. Therefore, when new cases are added or the sample set is changed, the previous clustering order is revised, and a completely new dendrogram is generated. This was observed when clustering was performed using the same methylation dataset (methylation at 449,005 CpG positions), however limiting the clustering to only ILBC cases (n=151). Additional Figure 1 in the appendices shows how the dendrogram changed after the clustering was limited to only ILBC cases. Samples from Subgroup 1 (26/28, 92%) (identified in all breast cancer clustering, Figure 4.1), clustered in branch A, separate from the rest of the ILBC cases, shown in black on the colour bar (Additional Figure 1). Samples from Subgroup 2 and Subgroup 3 clustered in the main branch B with some overlapping. Although it was reassuring that Subgroup 1 was recognised as the most distinct subgroup, the existence of further subgroups in addition to the three defined subgroups and the subjective nature of the sample assignment to the subgroups must be acknowledged. We used TCGA DNA methylation data of all breast cancer (n = 666) to replicate the clustering. A similar clustering pattern of ILBC was observed in the TCGA dataset with ILBC samples clustering beside the TNBC cases similar to the pattern observed in the study set (defined as Subgroup 1) (shown in black boxes in Additional Figure 2). Since the clustering of ILBC samples in TCGA dataset was more dispersed compared to the study set, we were unable to clearly assign samples to Subgroup 2 and Subgroup 3. This gives further evidence for the similarities of Subgroups 2 and 3 and further challenges the approach of distinguishing them as two distinct subgroups (as discussed in section 4.3.2). Differences

in the study set and TCGA dataset some of them being the characteristics of women such as age and tumour stage could likely be the cause of the observed differences in the clustering pattern. TCGA dataset, for example had on average older women with more advanced disease. We could not completely validate the three Subgroups in the TCGA dataset by performing the differential methylation analysis and linking the clusters to the clinical characteristics because it was challenging to assign samples to subgroups. This finding can be validated in larger dataset involving ILBC tumours with more uniform clinical characteristics to further assess if subgroups 2 and 3 are indeed one (albeit heterogeneous) subgroup. Another factor that could have impacted the methylation data is tumour purity. Tumour purity was assessed based on the methylation level information, which may be inaccurate without matching normal samples. No significant bias (*P*.value = 0.73, ANOVA) was observed between the subgroups in terms of the DNA methylation-based tumour purity estimate (Table 4.6). However, comparing these estimates with purity values generated by an experienced pathologist would be beneficial. Since tumour purity was not assessed as part of the routine histopathology review of the tumour material, this could not be achieved in this study.

4.5 Summary

Investigating the genome-wide tumour-derived DNA methylation status of ILBC tumours provided more evidence to support that there is substantial heterogeneity within this breast cancer subtype. This chapter provides data to support the hypothesis that subgroups of ILBC with increased homogeneity can be identified using genome-wide tumour DNA methylation measurements. As the somatic mutation profiles of ILBC methylation-defined subgroups could both provide some direct explanation for the methylation differences and some further evidence of heterogeneity (and thus potential

for defining more homogeneous ILBC subgroups) the next chapter explores somatic events associated with the three ILBC subgroups described in this chapter.

Chapter 5 Somatic Mutation Profiling of ILBC Methylation-defined Subgroups

5.1 Introduction

The molecular landscape of tumour is shaped by both genetic and epigenetic somatic events that are interlinked together (You & Jones, 2012). Growing evidence suggest that epigenetic modifications, in particular the DNA methylation pattern is influenced by the underlying genetic variations along with environmental and stochastic disruptions (Kerkel *et al.*, 2008; Bell *et al.*, 2011). DNA sequence variants that associate with DNA methylation are known as methylation quantitative trait loci (meQTL) and have been identified in different tissues throughout the genome (Gibbs *et al.*, 2010; Bell *et al.*, 2011; Drong *et al.*, 2013). The meQTLs are known to impact the DNA methylation pattern of local CpGs (*cis* acting) as well as CpGs located at distant sites (*trans* acting). Conversely, DNA methylation is also known to influence the tumour genome. Methylated cytosines are shown to be more prone to random deamination (loss of amino group) than unmethylated cytosines resulting in a higher rate of C > T mutations at methylated CpG positions in the genome (Ehrlich *et al.*, 1986). The rate of somatic mutation accumulation is also influenced by the nearby chromatic structure (among many other factors), which in turn is directly influenced by the methylation status of the DNA sequence (Schuster-Böckler & Lehner, 2012). Epigenetic silencing of DNA repair genes such as *MLH1*, *MGMT* and *BRCA1* has been reported to cause hypermutation during tumourigenesis that promotes genomic instability and enhance mutation rates in tumours (Toyota & Suzuki, 2010).

Unique combinations of mutation types in a cancer genome are a record of various endogeneous and exogeneous processes that have been active during tumour initiation and progression and are termed mutational signatures (Ludmil B Alexandrov & Stratton, 2014). Bioinformatics tools now make it possible to identify the contribution of these endogeneous and exogeneous processes in cancer genomes and thus can inform about the mechanisms that were disrupted during tumourigenesis (Ludmil B Alexandrov *et al.*, 2013). The identification of somatically acquired genomic variations provides the potential to identify driver mutations in tumours that may provide options for targeted therapies. The interpretation of mutational signatures provides additional opportunities to identify altered biological pathways that could be utilised in precision medicine approaches. For instance, the mutational signature, SBS3 has been associated with homologous recombination deficiency (HRD) (Nik-Zainal *et al.*, 2016). Pathogenic germline or somatic variation in *BRCA1* or *BRCA2* and promoter hypermethylation in *BRCA1* are well-recognised causes of HRD and thus, tumours showing SBS3 can potentially benefit from therapies targeting the HRD pathway (Nik-Zainal *et al.*, 2016). Large-scale analyses of 4,645 whole genome and 19,184 exome sequences from most types of cancers have currently defined 49 single base substitution (SBS), 11 doublet base substitution, four clustered base substitution and 17 small insertion and deletion signatures that have been attributed to different mutational processes (Ludmil B. Alexandrov *et al.*, 2020).

Chapter 4 described three subgroups of ILBC based on an unsupervised cluster analysis of the genome-wide DNA methylation data presented in section 4.3.1. Further analyses suggested that Subgroup 1 had a distinct methylation profile with a higher frequency of hypomethylation across the DMPs compared with the other two subgroups.

Subgroup 3 on the other hand, was found to show higher frequency of hypermethylation across the DMPs. Genome-wide differences in DNA methylation in the ILBC methylation-defined subgroups might be associated with somatic variations that are reflected in the mutational signatures. In this chapter, the whole exomes of ILBC tumours from the methylation-defined subgroups were investigated and identified the mutational signatures with an aim to further characterise the subgroups based on their somatic mutation profiles and identify any subgroup-specific mutational processes.

5.2 Method overview

5.2.1 Study participants and data

Analyses in this chapter included a total of 17 ILBC cases. Details of the study resources are provided in Table 5.1.

Table 5.1: Details of the samples and data being used in this chapter.

Result	Part I: Pilot study	Part II: Whole-exome sequencing and mutational signatures of tumours in the ILBC methylation-defined subgroups
Case	ILBC (n=2)	ILBC (n=15)
Sample type	Tumour - DNA derived from FFPE tissue Germline - DNA derived from GC and WB (Details of sample preparation are presented in the Methods sections 2.3.1 and 2.3.2).	Tumour - DNA derived from FFPE tissue Germline - DNA derived from GC
Total	Tumour (n=4) (Sample 1-FFPE, Sample 1-FFPE-rep†, Sample 2-FFPE, Sample 2-FFPE-rep) Germline (n=4) Sample 1-WB, Sample 1-GC Sample 2-WB, Sample 2-GC	Tumour (n=15) Germline (n=15)
Study resource	ABCFR[a] Details of the study design are presented in the Methods sections 2.1.1 and 2.1.3.	MCCS[b]
Data information	Somatic WES data (library preparation using Agilent SureSelect XT low input library preparation kit and SureSelect CREv2 exome (Agilent) as capture. (Details of library preparation and sequencing are presented in the Methods section 2.10).	

ILBC: Invasive lobular breast cancer. FFPE: Formalin-fixed paraffin-embedded. † Technical replicate are aliquots of DNA extraction from the same tumour sample. GC: Dried blood spots stored on Guthrie Cards. WB: Frozen whole blood. [a] Australian Breast Cancer Family Registry. [b] Melbourne Collaborative Cohort Study. WES: Whole-exome sequencing. CREv2: Clinical Research Exome v2.

Part I: Pilot study

5.3 Methods specific to part I

5.3.1 Experiment design

A pilot study was conducted to assess: i) the effect of input FFPE-derived DNA and genomic DNA quantity and quality on the sequencing data quality and ii) the suitability of DNA derived from dried blood spots stored on Guthrie Cards, referred to as GC-DNA, as an alternative to DNA derived from frozen whole blood, referred to as WB-DNA, as the matching germline sample. The pilot study data was also evaluated to determine the amount of sequencing data required for somatic variant calling and mutational signature analysis.

The pilot study included two ILBC cases as shown in Figure 5.1. ILBC case 2 was a known carrier of a heterozygous germline variant in *BRCA2* (NM_000059.3:c.8167G>C), which is classified as pathogenic by ENIGMA (Spurdle *et al.*, 2012). ILBC case 1 was not known to carry any pathogenic germline variant. As shown in Figure 5.1, each ILBC case had two tumour samples, which were DNA templates prepared from FFPE tumour material, referred to as Sample 1-FFPE and Sample 1-FFPE-rep for case 1 and Sample 2-FFPE and Sample 2-FFPE-rep for case 2. The “rep” samples were technical replicates, i.e., aliquots of DNA extraction from the same tumour from each ILBC case. Germline samples were prepared for each ILBC case from two different DNA sources, one from WB-DNA referred to as Sample 1-WB and Sample 2-WB and the other from GC-DNA referred to as Sample 1-GC and Sample 2-GC for case 1 and case 2, respectively.

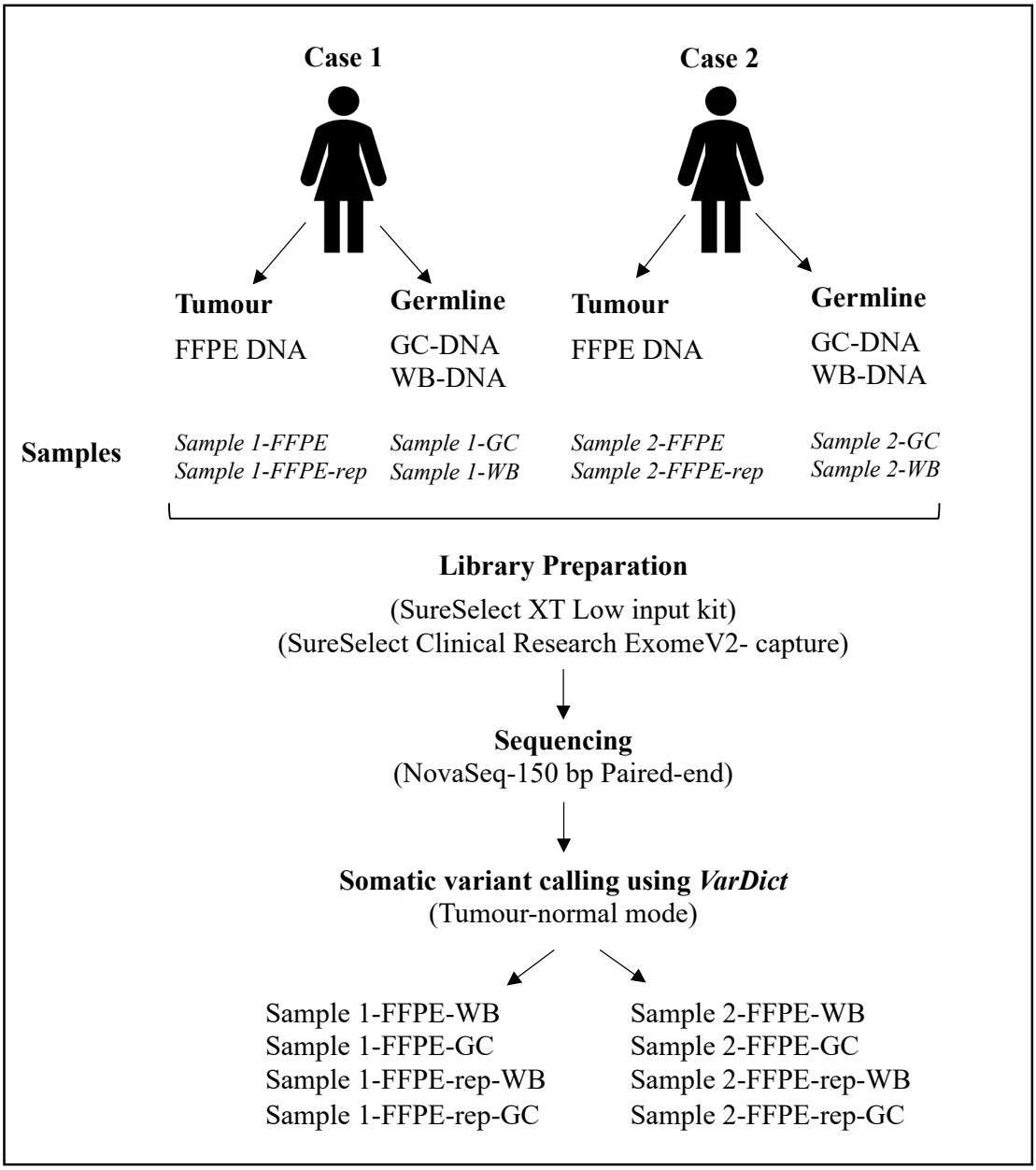


Figure 5.1: Selection of ILBC cases and library preparation for the pilot study.

Diagram showing an overview of the pilot study workflow. Libraries were prepared for two ILBC cases from DNA derived from formalin-fixed paraffin-embedded (FFPE) tumour material. Matching germline libraries were prepared from DNA derived from frozen whole blood (WB-DNA) and dried blood spots stored on Guthrie Cards (GC-DNA). Technical replicates of tumour, which were aliquots of DNA extraction from the same tumour are indicated as "rep". SureSelect XT low input kit (Agilent) was used for library preparation and the libraries were enriched using SureSelect Clinical Research Exome v2 (CREv2, Agilent). The libraries were pooled and sequenced on a S4 flow cell on NovaSeq to generate 150 bp paired end reads. Somatic variant calling was performed in paired tumour-normal analysis mode using *VarDict* (Lai *et al.*, 2016) using WB-DNA and GC-DNA as the reference germline.

DNA quality, as assessed using the DNA integrity score (ΔΔCq), and amplifiable DNA quantity were estimated for both the tumour and germline DNA samples for each case using the NGS FFPE qPCR QC assay (Agilent) as described in section 2.10.1 of the thesis. The input DNA quantity for library preparation was determined based on the ΔΔCq score as per the manufacturer's instructions. For DNA samples with a ΔΔCq > 1, input DNA quantity was calculated based on the qPCR assay results, whereas for samples with a ΔΔCq < 1, input DNA quantity was calculated based on the dsHS Qubit assay results (Thermo Fisher Scientific). An input DNA quantity of 100 ng was used for library preparation for all the samples except for Sample 2-FFPE-rep. An input DNA of 200 ng was used for Sample 2-FFPE-rep to test the effect of input DNA quantity on the sequencing data quality. Table 5.2 summarises the input DNA quantity and quality of the tumour and germline DNA samples in the pilot study.

Table 5.2: Input DNA quantity and quality of the samples in the pilot study.

Sample	Input DNA quantity (ng)	Input DNA quality (ΔΔCq)
Tumour		
Sample 1-FFPE	100*	1.19
Sample 1-FFPE-rep	100*	1.18
Sample 2-FFPE	100*	1.45
Sample 2-FFPE-rep	200*	1.3
Germline		
Sample 1-WB	100†	-0.11
Sample 1-GC	100†	0.31
Sample 2-WB	100†	-0.57
Sample 2-GC	100†	-0.08

ΔΔCq: DNA integrity score measured based on the NGS qPCR QC assay (Agilent). For samples with ΔΔCq > 1, input DNA quantity was calculated based on the NGS qPCR QC assay and for samples with ΔΔCq < 1, input DNA quantity was calculated based on the dsHS Qubit assay. * Input DNA amount calculated based on the NGS qPCR QC assay. † Input DNA amount calculated based on the Qubit assay. The replicate samples indicated as "rep" are aliquots of DNA extraction from the same tumour. ng: nanogram.

Libraries were prepared using the SureSelect XT low input library preparation kit (Agilent) and enriched using Agilent SureSelect CREv2 as described in section 2.10.3 of the thesis. Tumour and germline libraries were pooled for multiplex sequencing as described in section 2.10.4 of the thesis. 2.5 nM of tumour libraries and 2.5 nM of germline libraries were pooled in a 4:1 (tumour: germline) ratio with the aim to achieve a mean target depth of coverage for tumour and germline samples of 150X and 50X, respectively. Sequencing was performed on the NovaSeq 6000 (Illumina, USA) using an S4 flow cell at the Australian Genome Research Facility (AGRF). All the libraries (n=8) were sequenced on a single lane to generate 150 bp paired end reads.

5.3.2 Sequencing data processing and somatic variant calling

Raw sequencing data was pre-processed as described in section 2.11 of the thesis. Somatic variant calling was performed using *VarDict* (Lai *et al.*, 2016) for tumour samples and their technical replicates in paired tumour-normal analysis mode, using WB-DNA and GC-DNC as the reference germline and two sets of somatic variant calls were generated. Somatic variant calling was performed using *VarDict* default filters as presented in detail in section 2.11 of the thesis.

5.3.3 Determining thresholds for somatic variant filtering

Single nucleotide variants (SNVs) that were exclusively detected in the tumour and not in the germline DNA (tagged as “StrongSomatic” by *VarDict*) and had passed all the default *VarDict* filters (tagged as “PASS”) were considered somatic SNVs (SSNVs) and selected for further analyses. A detailed description of the *VarDict* default filters is presented in section 2.11 of the thesis. On the SSNVs the following additional filters were applied: i) minimum read depth of 30X and ii) minimum variant allele fraction of 0.2. These filters were selected considering the mean target depth of coverage and the

percentage of target coverage achieved across the samples. For determining the above filter cut-offs, the BAM files of tumour and the reference germline samples were compared. The concordance level (number of matching genotypes) between the tumour and normal reads were assessed using the *bcftools* (H Li *et al.*, 2009). The tumour and normal reads were compared at different allele frequencies and depths considering that a lower concordance level mainly represents the presence of sequencing artefacts.

5.3.4 Variant annotation and mutational signature analysis

The SSNVs were annotated using the *VarSeq* software (Golden Helix, Inc., Bozeman, MT, www.goldenhelix.com), to add annotations to the somatic and germline variants that included *RefSeq Genes* (O'Leary *et al.*, 2016), *COSMIC Mutations* (Tate *et al.*, 2019) and *ClinVar* (Landrum *et al.*, 2018).

Mutational signatures were generated using the R package *deconstructSigs* (Rosenthal *et al.*, 2016). The somatic signature profiles of the tumour samples were identified using the predefined mutational signatures COSMIC (version3) (http://cancer.sanger.ac.uk/cosmic/signatures). Mutational signatures with weight 0.06 and higher were considered significant as described in *deconstructSigs* (Rosenthal *et al.*, 2016).

5.3.5 Determining the concordance between variant calls identified using different reference germline

The SSNVs identified in the tumour samples using GC-DNA and WB-DNA as the reference germline are henceforth referred to as $SSNVs^{GC}$ and $SSNVs^{WB}$, respectively. To determine the concordance between $SSNVs^{GC}$ and $SSNVs^{WB}$, the two

sets of somatic variant calls (.vcf files) were compared at each variant position using the *-isec* option of *bcftools* (H Li *et al.*, 2009). The function generated output .vcf files containing somatic variants that overlapped between $SSNVs^{GC}$ and $SSNVs^{WB}$ as well as .vcf files with somatic variants that were unique to $SSNVs^{GC}$ and $SSNVs^{WB}$ variant call sets.

5.4 Results: Part I

5.4.1 Evaluating the sequencing performance

In the pilot study, tumour exomes of two ILBC cases (Sample 1-FFPE and Sample 2-FFPE) were sequenced. Germline exomes from matching GC-DNA (Sample 1-GC and Sample 2-GC) and WB-DNA (Sample 1-WB and Sample 2-WB) were used as the reference germline. Technical replicates of the tumour samples (Sample 1-FFPE-rep and Sample 2-FFPE-rep) were also sequenced to test for technical reproducibility.

The sequencing yielded an average of 138 million unique reads and a mean target depth of 56X for the tumour samples. For germline GC-DNA samples, an average of 49 million unique reads and a mean target depth of 34X were achieved, whereas for germline WB-DNA samples an average of 48 million unique reads and a mean target depth of 37X were achieved (Figure 5.2a). For tumour samples, over 50% of the target was covered on average at a minimum of 30X read depth, whereas for the germline samples over 40% of the target was covered on average at a minimum of 30X read depth (Figure 5.2b). The mean target depth of coverage achieved in the pilot study was lower than expected for both the tumour (~150X) and germline (~50X) samples, as estimated based on the allocated data yield capacity of the NovaSeq S4 flow cell (120 Gb) and the targeted data

yield for the tumour and germline samples, which were 24 Gb and 6 Gb, respectively as recommended by AGRF. The sequencing quality metrics of the tumour and germline samples in the pilot study are summarised in Table 5.3.

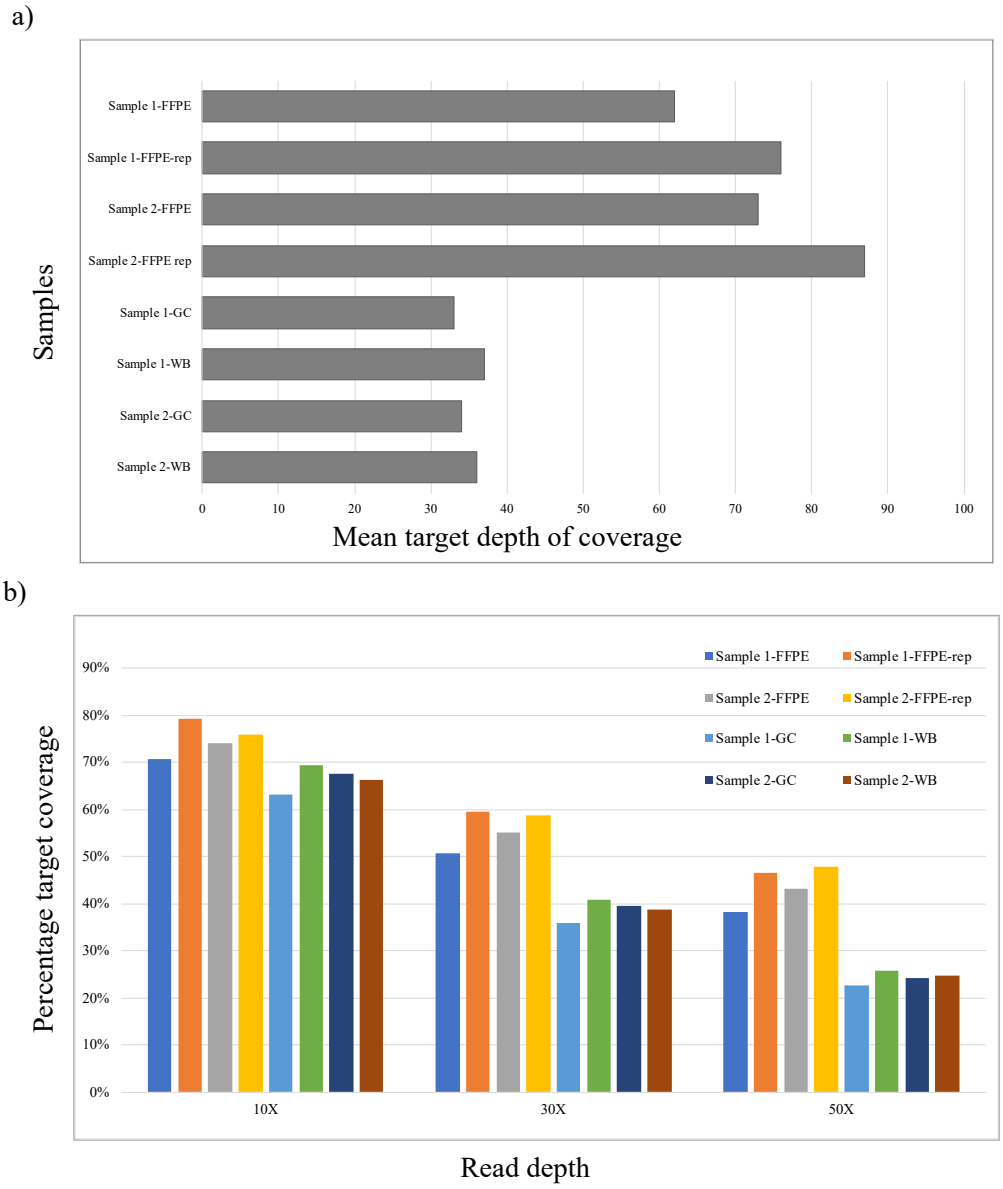


Figure 5.2: Mean target depth of coverage and percentage of target covered at 10X, 30X and 50X of the tumour and germline samples in the pilot study.

Graphs showing **a**) the mean target depth of coverage and **b**) the percentage of target (Clinical Research Exome v2, Agilent) covered at different read depths (10X, 30X and 50X) for the samples in the pilot study, calculated using *Picard* tools (http://broadinstitute.github.io/picard/). Samples in plot **b** are represented by different colour bars as indicated in the legend in the top-right.

Table 5.3: Whole-exome sequencing* quality metrics for the tumour and germline samples included in the pilot study.

Sample	Raw data yield (Gb)	Number of unique[a] reads (million)	Mean target* depth of coverage	Bases of target covered at least at 30X depth (%)	Bases of target not covered at all (%)	Bases off-target[†] (%)	PCR duplication rate (%)
Tumour							
Sample 1-FFPE	33	122	62	51	2.3	8	41
Sample 1-FFPE-rep	38	125	76	60	2	8	50
Sample 2-FFPE	33	137	73	55	2.2	6	36
Sample 2-FFPE-rep	37	169	87	59	2	6	29
Germline							
Sample 1-WB	10	48	37	41	3	7	21
Sample 1-GC	9	48	33	36	3	5	16
Sample 2-WB	10	48	36	39	3	6	20
Sample 2-GC	10	50	34	40	3	6	18

Gb: Gigabase. [a] Reads that are not marked as duplicates. *Clinical Research Exome v2, Agilent, 67.3Mb. [†] Bases that did not align the target region. Sequencing data quality metrics were calculated using *Picard* tools.

To assess if the input DNA quantity has an impact on the sequencing data quality, the sequencing performance of Sample 2-FFPE and Sample 2-FFPE-rep were compared, which used different input DNA quantities (100 ng and 200 ng, respectively). Sample 2-FFPE-rep had nominally higher mean target depth of coverage (87X *versus* 73X in Sample 2-FFPE) and percentage of target covered at a read depth of 30X (60% *versus* 55% in Sample 2-FFPE), and a lower PCR duplication rate (29% *versus* 36% in Sample 2-FFPE) in comparison with Sample 2-FFPE. The percentage of target covered at a read depth of 50X was also nominally higher for Sample 2-FFPE-rep (48% *versus* 43% in Sample 2-FFPE) (Figure 5.2b). No difference in the percentage of bases aligned off-target was observed between the two samples. The proportion of target bases not covered was also similar (~ 2%) for both samples (Table 5.3).

Input DNA quality, as assessed by the ΔΔCq score ranged from 1.18 to 1.45 across the tumour samples and from -0.57 to 0.31 across the germline samples (Table 5.2). A large variation in the ΔΔCq score was not observed across the same sample type, therefore a reliable comparison between DNA quality and the output sequencing data quality could not be made.

5.4.2 Use of Guthrie Card-derived DNA as germline reference

To assess the suitability of GC-DNA to be used as an alternative to WB-DNA as the germline reference sample in tumour-normal paired analysis, the SSNVs ($SSNVs^{GC}$ and $SSNVs^{WB}$) and the mutational signatures identified in the tumour samples including their technical replicates, using the two sources of germline DNA were compared.

At a minimum read depth of 30X, 33% of the target region was commonly covered by both Sample 1-WB and Sample 1-GC and 34% of the target region was commonly covered by both Sample 2-WB and Sample 2-GC (Table 5.4).

Table 5.4: Target base coverage for the two sources of germline DNA.

Sample	Bases of target* covered at least at 30X depth (%)	Number of target bases uniquely covered at a minimum read depth of 30X	Bases of target commonly covered by WB-DNA and GC-DNA samples at 30X depth (%)
Sample 1-WB	41	2,259	33
Sample 1-GC	36	11,444	
Sample 2-WB	39	9,414	34
Sample 2-GC	40	6,114	

* Clinical Research Exome v2 (Agilent, 67.3Mb). WB-DNA: DNA derived from frozen whole blood. GC-DNA: DNA derived from blood spots stored on Guthrie Cards.

The total number of $SSNVs^{GC}$ and $SSNVs^{WB}$ identified in Sample 1-FFPE, Sample 2-FFPE and their technical replicates are summarised in Table 5.5. For Sample 1-FFPE, a concordance level of over 55% between $SSNVs^{GC}$ and $SSNVs^{WB}$ was observed with 157 overlapping SSNVs, whereas 239 and 126 SSNVs were unique to the $SSNVs^{GC}$ and $SSNVs^{WB}$ call sets, respectively (Figure 5.3a). For Sample 2-FFPE, a concordance level of 55% was observed with 215 overlapping SSNVs, whereas 174 and 182 SSNVs were unique to the $SSNVs^{GC}$ and $SSNVs^{WB}$ call sets, respectively (Figure 5.3b). For Sample 1-FFPE-rep, a lower concordance level of 33% was observed with 199 overlapping SSNVs, whereas 407 and 219 SSNVs were unique to the $SSNVs^{GC}$ and $SSNVs^{WB}$ call sets, respectively (Figure 5.3c). For Sample 2-FFPE-rep, a concordance level of over 50% was observed with 279 overlapping SSNVs, whereas 223 and 295 SSNVs were unique to the $SSNVs^{GC}$ and $SSNVs^{WB}$ call sets, respectively (Figure 5.3d).

Table 5.5: Total number of somatic single nucleotide variants identified in the tumour samples and their technical replicates using DNA derived from frozen whole blood and DNA derived from dried blood spots stored on Guthrie Cards as the germline reference sample.

	WB-DNA (germline reference)	**GC-DNA (germline reference)**
Tumour	**SSNVs**	**SSNVs**
Sample 1-FFPE	283	396
Sample 1-FFPE-rep	418	606
Sample 2-FFPE	397	389
Sample 2-FFPE-rep	574	502

WB-DNA: DNA derived from frozen whole blood. GC-DNA: DNA derived from dried blood spots stored on Guthrie Cards. FFPE: Formalin-fixed paraffin-embedded. SSNV: Somatic single nucleotide variant.

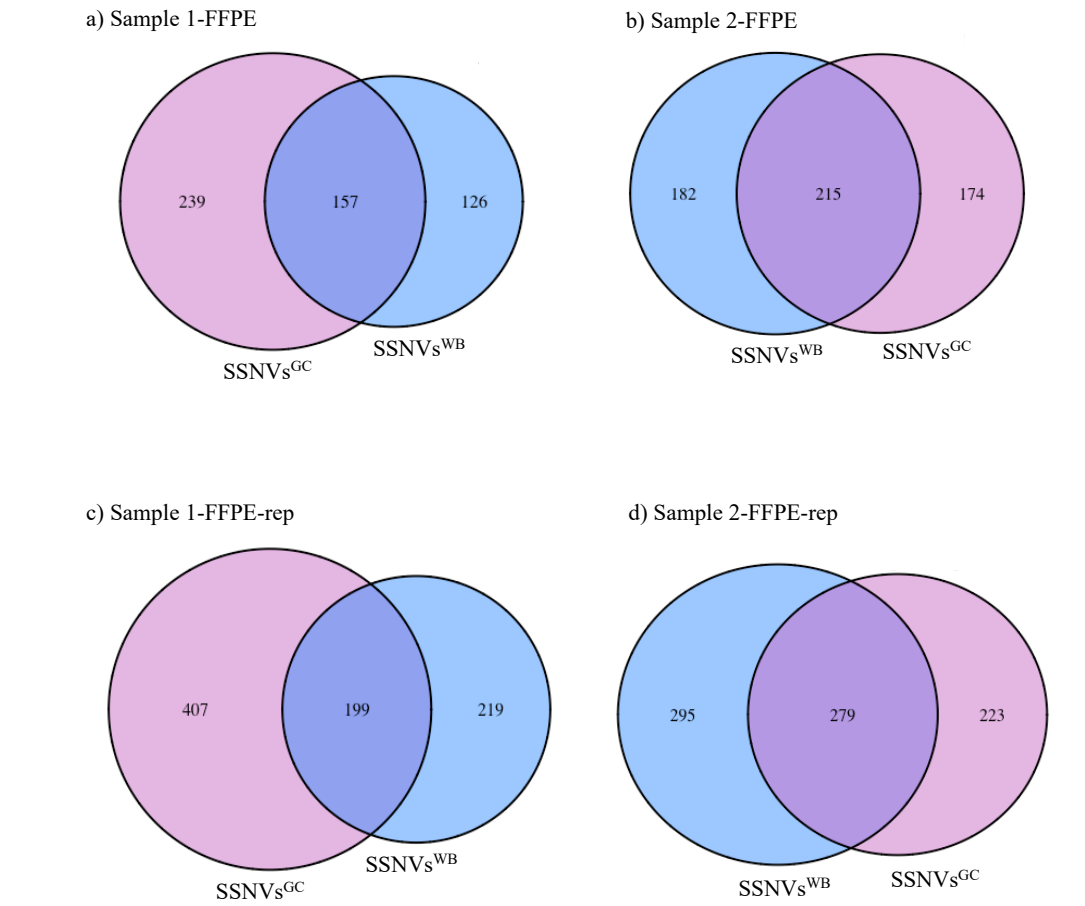


Figure 5.3: Comparison of somatic single nucleotide variants calls identified using DNA derived from frozen whole blood and DNA derived from dried blood spots stored on Guthrie Cards as germline reference.

Venn diagrams showing the overlap between somatic single nucleotide variants (SSNVs) identified in the tumour samples and their technical replicates; **a**) Sample 1-FFPE, **b**) Sample 2-FFPE, **c**) Sample 1-FFPE-rep and **d**) Sample 2-FFPE-rep, using matching DNA derived from frozen whole blood sample (WB-DNA) and dried blood spots stored on Guthrie Cards (GC-DNA) as the germline reference sample, in the tumour-normal paired analysis, referred to as $SSNVs^{WB}$ and $SSNVs^{GC}$, respectively.

To investigate if the source of germline DNA has any impact on mutational signature analysis, mutational signatures were generated using the two sets of variant calls, $SSNVs^{WB}$ and $SSNVs^{GC}$. Some mutational signatures were detected from both the pairs, whereas some signatures were detected from only one of the pairs. For Sample 1-FFPE, mutational signatures SBS4 (signature weight = 0.06), SBS6 (0.06) and SBS31 (0.11) were identified only from $SSNVs^{GC}$, whereas SBS16 (0.08) was identified from $SSNVs^{WB}$ and not from $SSNVs^{GC}$. Sample 1-FFPE-rep on the other hand, showed more concordant signature profile with only two signatures, SBS31 (0.07) and sequencing artefacts associated signatures (0.07), uniquely identified from $SSNVs^{GC}$ and not from $SSNVs^{WB}$ (Figure 5.4).

In the case of Sample 2-FFPE, with a known heterozygous germline pathogenic variant in *BRCA2* (NM_000059.3:c.8167G>C), the expected HRD-associated signature (Mesman *et al.*, 2019), SBS3 was identified from $SSNVs^{GC}$ as the most predominant signature (signature weight=0.26), however $SSNVs^{WB}$ did not identify this signature (Figure 5.4). Other signatures that were uniquely identified from $SSNVs^{GC}$ were SBS5 (0.07), SBS6 (0.08) and SBS13 (0.06). Mutational signatures identified uniquely from $SSNVs^{WB}$ were SBS1 (0.07), SBS5 (0.31), SBS24 (0.08) and SBS39 (0.08). Sample 2-FFPE-rep displayed SBS3 in both the pairs with a signature weight of 0.20 and 0.15, respectively (Figure 5.4). No discrepancies were observed in the mutational signatures identified from $SSNVs^{GC}$ and $SSNVs^{WB}$ for Sample 2-FFPE-rep, although there were some differences in signature weight.

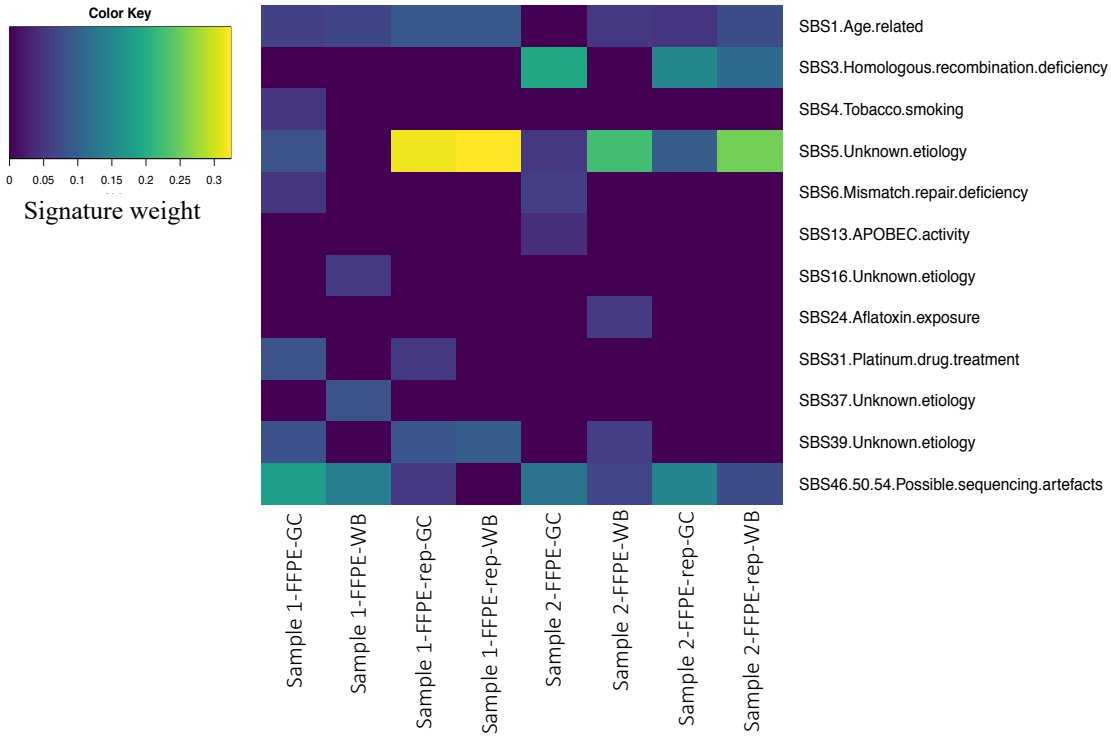


Figure 5.4: Mutational signatures identified in the tumour of ILBC cases included in the pilot study.

Heatmap showing the mutational signatures identified in the tumour of ILBC cases generated using the somatic single nucleotide variants (SSNVs) with DNA derived from frozen whole blood (WB-DNA) and DNA derived from dried blood spots stored on Guthrie Cards (GC-DNA) used as matching germline reference sample in a tumour-normal paired analysis. The mutational signatures identified using the SSNVs with GC-DNA as the germline reference are marked as "FFPE-GC" and the mutational signatures generated using the SSNVs with WB-DNA as the germline reference are marked as "FFPE-WB" on the heatmap. Mutational signatures were generated using *deconstructSigs* (Rosenthal *et al.*, 2016). The samples are plotted on the x-axis (as columns) and the mutational signatures are plotted on the y-axis (as rows). Signature weights are represented by colours in the heatmap as indicated in the colour key in the left corner.

5.4.3 Concordance between tumour samples and their replicates

Tumour replicate libraries were prepared from aliquots of DNA extracted from the same tumour samples. The library profiles and SSNVs of the tumour samples were compared to their replicates to test for technical reproducibility of the assay.

Comparing the post-hybridisation library profiles of the tumour samples and their respective technical replicates, Sample 1-FFPE library showed a fragment size peak at 282 bp (Figure 5.5a) and the profile resembled a typical FFPE DNA sample as recommended in the protocol with the expected peak size of 200-400 bp, whereas in the case of Sample 1-FFPE-rep, two peaks were observed, one at 279 bp and another one at 324 bp (Figure 5.5b). In the case of Sample 2-FFPE and its replicate, DNA fragment peaks were observed within the expected range at 284 bp and 274 bp, respectively (Figure 5.6a and Figure 5.6b).

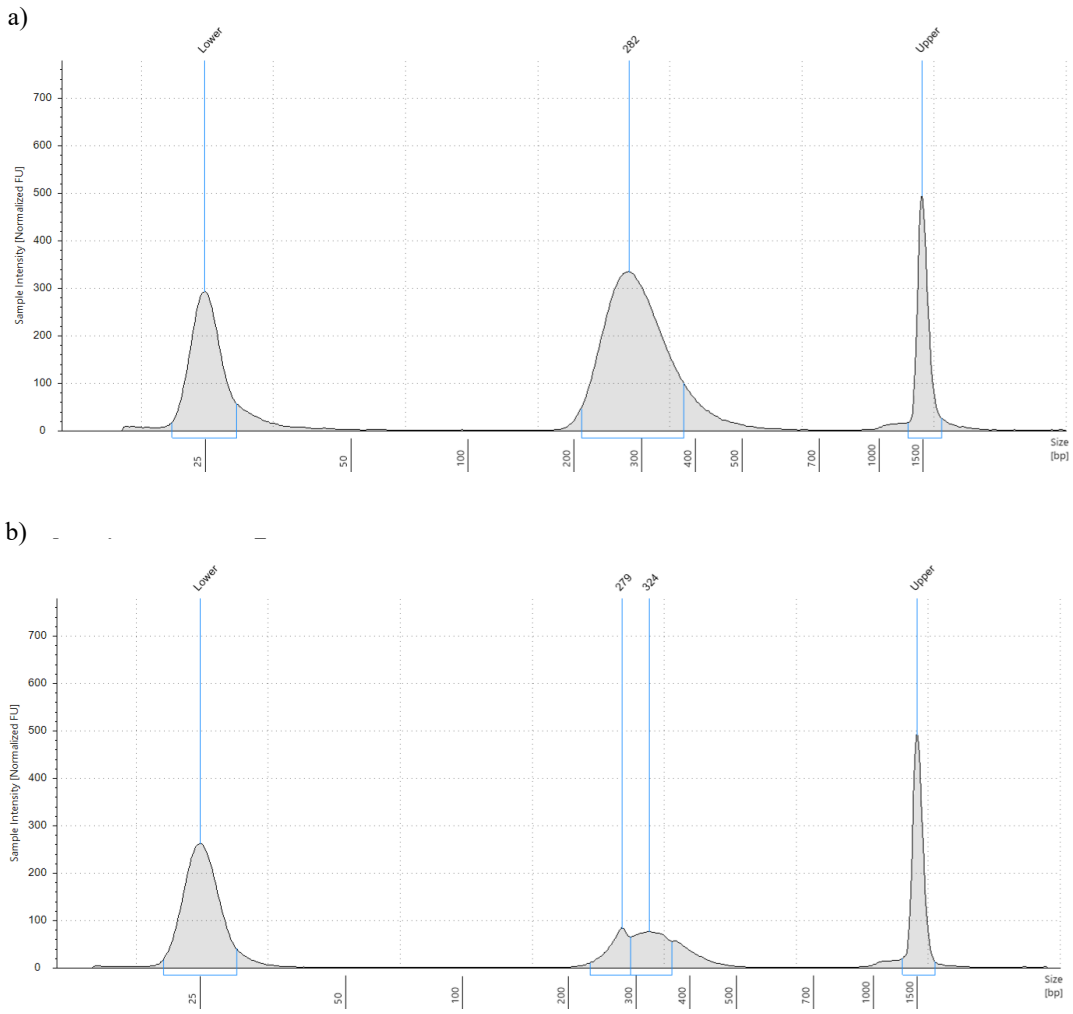


Figure 5.5: Post-hybridisation library profiles of Sample 1-FFPE and its technical replicate.

Electropherograms showing the post-hybridisation library profiles of **a**) Sample 1-FFPE and **b**) Sample 1-FFPE-rep, separated using the D1000 ScreenTape assay and the TapeStation system. The fragment size distribution of the library is shown on the x-axis and the library intensity is shown on the y-axis. The lower and upper markers at 25 bp and 1500 bp are internal references that are used to determine the molecular weight size of the sample. The expected peak size of the library DNA fragment is 200-400 bp according to the Agilent SureSelect low input library preparation protocol. The peak size of Sample 1-FFPE library was 282 bp. Two library fragment peaks were observed for Sample 1-FFPE-rep, one at 279 bp and another at 324 bp.

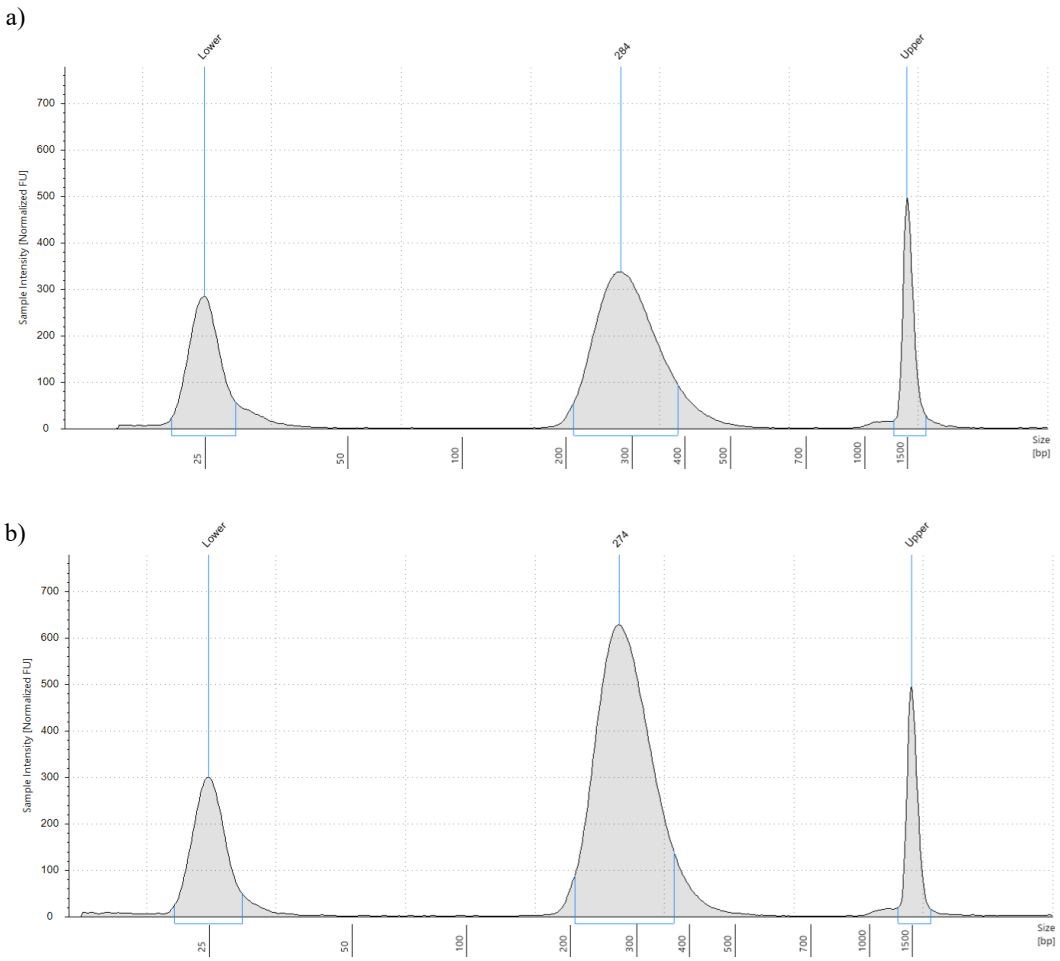


Figure 5.6: Post-hybridisation library profiles of Sample 2-FFPE and its technical replicate.

Electropherograms showing the post-hybridisation library profiles of **a**) Sample 2-FFPE and **b**) Sample 2-FFPE-rep, separated using the D1000 ScreenTape assay and the TapeStation system. The fragment size distribution of the library is shown on the x-axis and the sample intensity is shown on the y-axis. The lower and upper markers at 25 bp and 1500 bp are internal references that are used to determine the molecular weight size of the sample. The expected peak size of the library DNA fragment is 200-400 bp according to the Agilent SureSelect low input library preparation protocol. The peak size of Sample 2-FFPE and Sample 2-FFPE-rep libraries was observed at 284 bp and 274 bp, respectively. The tumour sample and its replicate differed in their input DNA amount where 100 ng and 200 ng of input DNA was used for Sample 2-FFPE and Sample 2-FFPE-rep, respectively.

At a minimum read depth of 30X, 59% of the target bases were commonly covered by both Sample 1-FFPE and Sample 1-FFPE-rep, whereas 65% of the target bases were commonly covered by both Sample 2-FFPE and Sample 2-FFPE-rep. Similar SSNVs call sets were expected in the tumour samples and their respective technical replicates given that they were aliquots of DNA extraction from the same tumour sample. Comparing the $SSNVs^{WB}$ variant call set, Sample 1-FFPE and Sample 1-FFPE-rep showed a concordance of 41%, whereas a concordance of 64% was observed between Sample 2-FFPE and Sample 2-FFPE-rep (Figure 5.7a, Figure 5.7c). On the other hand, comparing the $SSNVs^{GC}$ variant call set, Sample 1-FFPE and Sample 1-FFPE-rep showed a concordance of 53%, whereas a concordance of 58% was observed between Sample 2-FFPE and Sample 2-FFPE-rep (Figure 5.7b, Figure 5.7d).

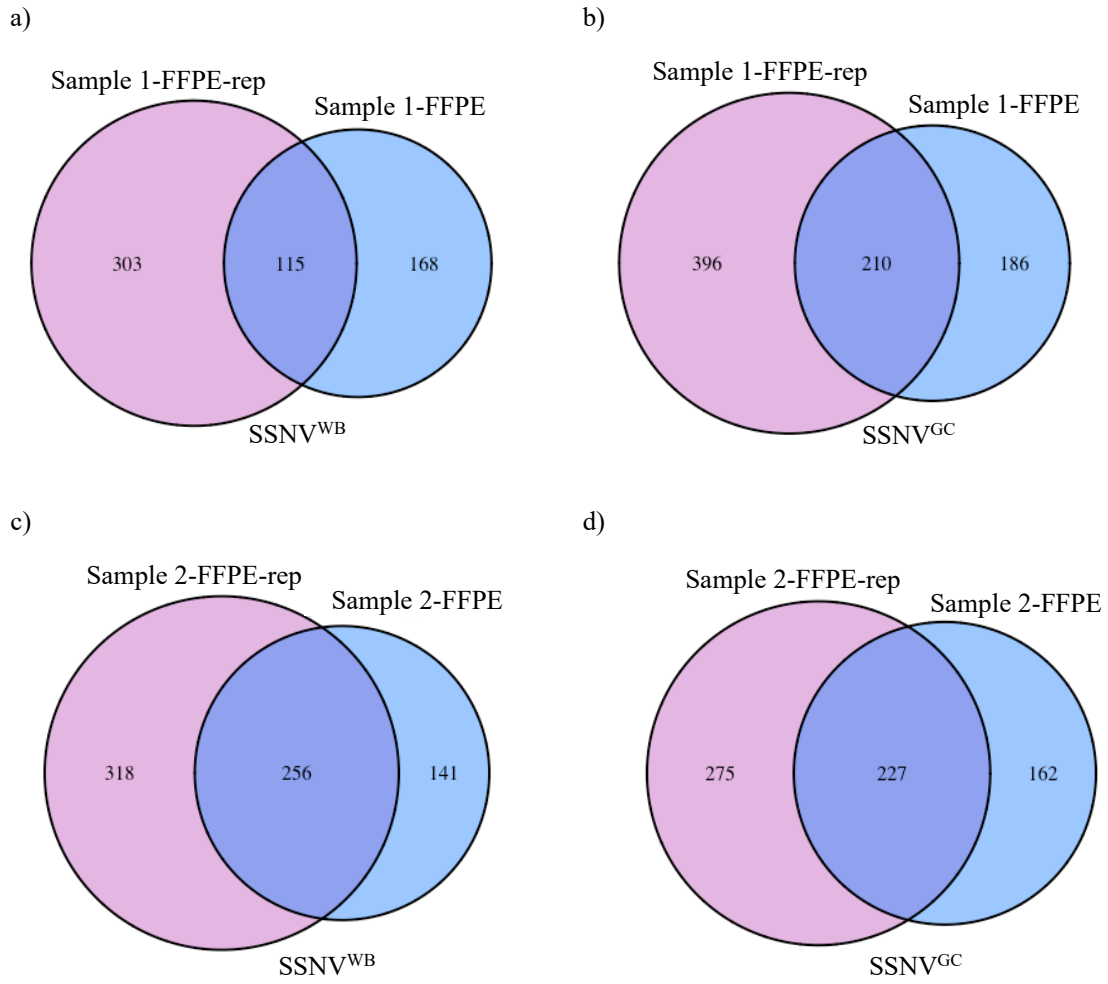


Figure 5.7: Comparison of somatic single nucleotide variants identified in the tumour samples and their technical replicates.

Venn diagrams showing the overlap of somatic single nucleotide variants (SSNVs) identified in the tumour samples and their respective technical replicates. The SSNVs identified using DNA derived from frozen whole blood (WB-DNA) as the germline reference sample are referred to as $SSNVs^{WB}$ and the SSNVs identified using DNA derived from dried blood spots stored on Guthrie Cards (GC-DNA) as the germline reference are referred to as $SSNVs^{GC}$. Venn diagrams **a**) and **b**) show the comparison between Sample 1-FFPE and Sample 1-FFPE-rep of $SSNVs^{WB}$ and $SSNVs^{GC}$ variant call sets, respectively. Venn diagrams **c**) and **d**) show the comparison between Sample 2-FFPE and Sample 2-FFPE-rep of $SSNVs^{WB}$ and $SSNVs^{GC}$ variant call sets, respectively.

Part II: Whole-exome sequencing and mutational signatures of tumours in the ILBC methylation-defined subgroups

5.5 Methods specific to part II

5.5.1 Sample selection

Five ILBC cases were selected from each ILBC methylation-defined subgroup identified in Chapter 4 for WES (Figure 5.8). Table 5.6 summarises the clinical and pathological characteristics of the selected ILBC cases. The primary selection criteria for the cases from the subgroups was their participation in the MCCS. Samples were selected only from the MCCS to minimise the differences in study design and data collection methods.

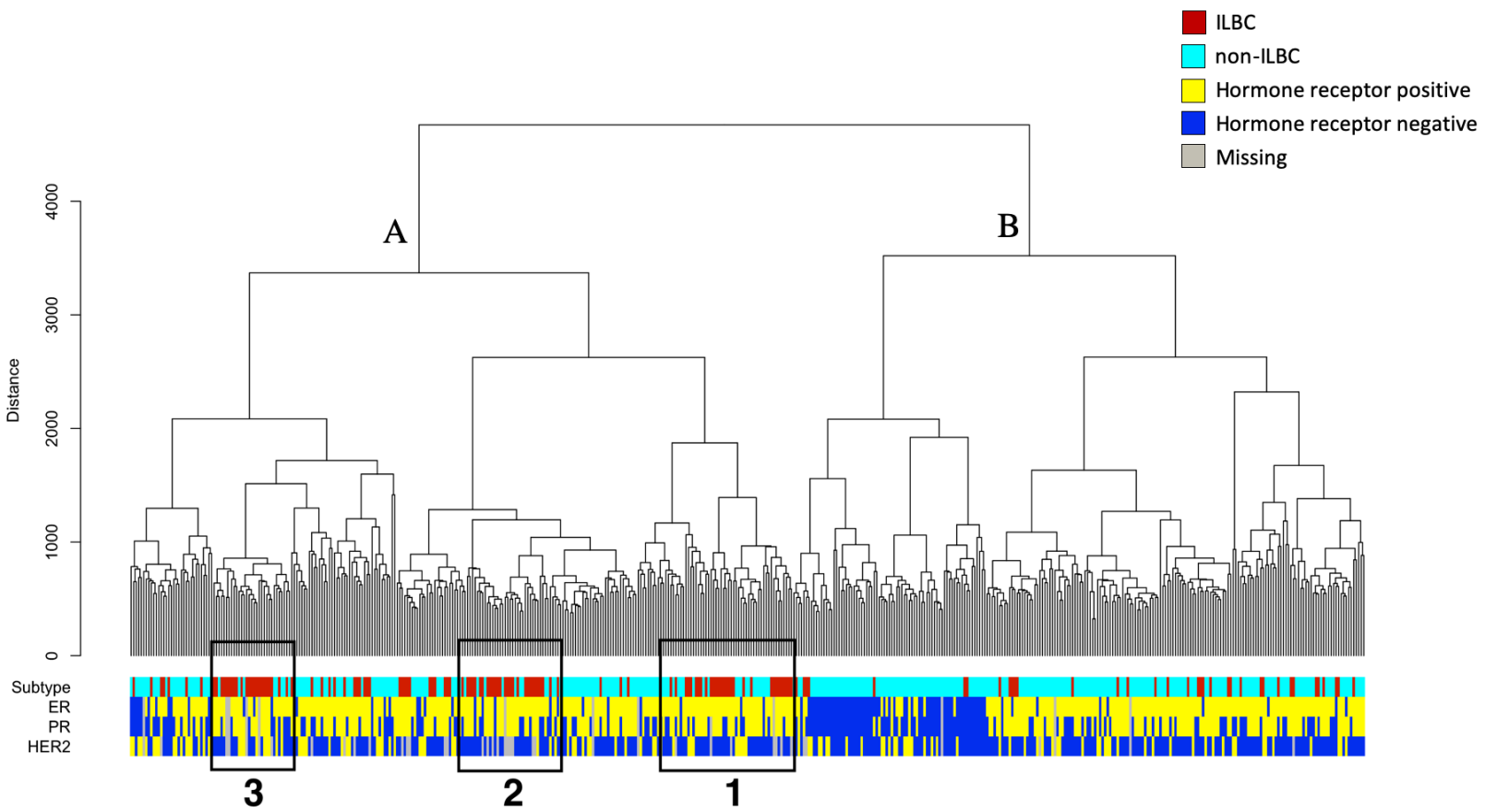


Figure 5.8: Selection of cases from the ILBC methylation-defined subgroups for whole-exome sequencing.

Dendrogram showing the result of the cluster analysis presented in Chapter 4 where three methylation-defined subgroups were identified (shown in black boxes). The colour bar "Subtype" shows ILBC samples in red and non-ILBC samples in turquoise colour. Five cases were selected from each ILBC methylation-defined subgroup for whole-exome sequencing.

Table 5.6: Clinical and pathological characteristics of the ILBC cases from the methylation-defined subgroups selected for whole-exome sequencing.

Case	ILBC subgroup	Age at diagnosis	ICDO-code	Smoking status	Tumour grade	Tumour stage	Hormone receptor status	Female relative with breast cancer*	Family history of colorectal cancer
8965	1	70	8520	Never	II	1A	ER+/PR+/HER2-	NA	NA
34757	1	83	8520	Never	II	2A	ER+/PR+/HER2-	Yes (Mother)	No
13412	1	67	8520	Never	II	1A	ER+/PR-/HER2-	No	No
10780	1	62	8522	Never	II	3A	ER+/PR+/HER2+	No	NA
27328	1	57	8520	Never	II	2A	ER+/PR+/HER2-	No	No
34778	2	67	8520	Never	II	1A	ER-/PR-/NA	Yes	Yes
24634	2	60	8520	Never	II	3A	ER+/PR+/HER2-	No	No
29513	2	71	8520	Never	II	1A	ER+/PR-/HER2-	Yes (Aunt)	Yes
27217	2	69	8520	Former	II	1A	ER+/PR+/HER2-	NA	No
38184	2	63	8520	Never	III	1A	ER+/PR+/HER2-	No	Yes
33053	3	69	8522	Never	II	3A	ER+/PR+/HER2-	No	Yes
25911	3	68	8520	Never	II	1A	ER+/PR+/HER2-	Yes (Mother)	No
40384	3	51	8522	Former	II	2A	ER+/PR-/HER2-	No	No
37762	3	51	8522	Current	III	1A	ER+/PR+/HER2+	Yes (Mother)	No
32623	3	60	8520	Never	II	2A	ER+/PR+/HER2-	Yes (Aunt)	Yes

ER: Estrogen. PR: Progesterone. HER2: Human epidermal growth factor receptor 2, +: Positive. -: Negative. NA: Not available. *Any breast cancer type. ICDO: International Classification of Diseases for Oncology (8520- Infiltrating lobular carcinoma, 8522- Infiltrating ductal and lobular carcinoma).

5.5.2 Library preparation and sequencing

Tumour and germline libraries were prepared for 15 ILBC cases from DNA derived from tumour-enriched FFPE material and GC-DNA, respectively. Library preparation was done following the manufacturer's instructions without any modifications, as described in section 2.10.3 of the thesis. An input DNA quantity of 200 ng (where available) was used for library preparation for both the tumour and germline samples as per the manufacturer's recommendation. Based on the comparison of WB-DNA and GC-DNA in the pilot study, GC-DNA was found suitable to be used as an alternative to WB-DNA as the germline reference and, therefore all germline libraries in the main experiment were prepared using GC-DNA. As the mean target depth of coverage achieved in the pilot study was lower than expected, additional sequencing was added in the main experiment to mitigate the loss of sequencing due to off-target coverage and PCR duplication. We aimed for a mean target depth of 150X mapped for tumour and 40X-50X mapped for the germline samples. The data output for the main experiment was estimated based on the pilot study data. In the pilot study for tumour samples,

$$35\ \text{Gb raw data} \rightarrow 138\ \text{million reads average} \rightarrow 75\text{X mean target depth of coverage}$$

So, to achieve 150X mapped mean target depth of coverage for the tumour samples,

$$\text{Required raw data} = 35 * \frac{150}{75} = 70\ \text{Gb}$$

For germline samples,

$$10\ \text{Gb raw data} \rightarrow 49\ \text{million reads average} \rightarrow 35\text{X mean target depth of coverage}$$

So, to achieve 40X-50X mapped mean target depth of coverage for the germline samples,

$$\text{Required raw data} = 10 * \frac{40}{35} = 10\ \text{Gb}$$

Although the calculations indicated 70 Gb of raw data per tumour sample and 10 Gb of raw data per germline sample, the targeted data was considered too high by AGRF. Considering the low library complexity of tumour samples, increasing the sequencing after a certain point would not result in increase in the target depth of coverage as the same DNA template are sequenced rather than sequencing any more unique DNA template. So, considering AGRF's suggestion, we aimed for 44 Gb of raw data per tumour sample and 11 Gb of raw data per germline sample.

5.5.3 Tumour mutation burden

Tumour mutation burden (TMB) was calculated as the total number of SSNVs per megabase of the target region covered. Since the percentage of target region covered at a read depth of 30X (minimum read depth used for variant filtering) was different for different tumour samples, the effective target region covered of the total 67.3 Mb (target capture size-CREv2) for each sample was calculated as below and individual target coverage value for each sample was used in the TMB calculation.

$$\text{Effective target region} = \%\ \text{target covered at a read depth of } 30\text{X}/67.3\ \text{Mb}$$

5.5.4 Testing for microsatellite instability

Microsatellite instability of the tumour samples were investigated using *MSIsensor* (Niu *et al.*, 2014). The reference genome was first interrogated for microsatellites (maximum repeat unit length of 5 bp) and homopolymers (at least 5 bp length) and the microsatellite sites found in the reference genome were recorded. Aligned sequencing reads of both tumour and germline (BAM files) with sufficient coverage (at leads 20 reads in both the tumour and germline) were then examined for the available microsatellite regions, recorded previously in the reference genome and deletion length variation between tumour and germline was identified. A χ^2 test was used to identify the

loci that were significantly different between tumour and germline and were tagged as somatic. The percentage of somatic sites identified based on the χ^2 test were represented as the MSI score. The samples with MSI score > 3.5 were considered microsatellite instable and with MSI score < 3.5 were considered microsatellite stable as described in (Niu *et al.*, 2014).

5.6 Results: Part II

5.6.1 Evaluating the sequencing performance

To investigate the mutational signatures in the tumours from the three ILBC methylation-defined subgroups, WES was performed on five cases from each subgroup using DNA derived from FFPE tumour material and matching germline DNA derived from GC-DNA (Figure 5.8). The input DNA quality and quantity and the sequencing data metrics of the tumour samples and the germline reference samples are summarised in Table 5.7 and Table 5.8, respectively.

Table 5.7: Input DNA quality and quantity and whole-exome sequencing* data quality metrics of tumour samples from the ILBC methylation-defined subgroups calculated using *Picard*.

Tumour sample (FFPE)	ILBC subgroup	Input DNA quantity (ng)	Input DNA quality (ΔΔCq)	Number of unique[a] reads (million)	Mean target* depth of coverage	Bases of target covered at least at 30X (%)	Bases of target not covered at all (%)	Bases off-target[†] (%)	PCR duplication (%)
8965_FFPE	1	200	3.1	149	57	48	2	11	60
34757_FFPE	1	200	2.8	218	106	59	2	6	56
13412_FFPE	1	177	4.6	263	116	61	2	7	70
10780_FFPE	1	200	4.0	119	38	37	2	10	58
27328_FFPE	1	200	3.3	188	81	51	2	8	58
34778_FFPE	2	200	0.83	246	158	59	2	5	54
24634_FFPE	2	200	1.3	175	80	61	2	10	49
29513_FFPE	2	179	3.7	177	83	57	2	8	57
27217_FFPE	2	200	1.76	217	107	58	2	5	50
38184_FFPE	2	200	1.78	214	93	52	2	7	52
33053_FFPE	3	200	0.99	273	160	65	2	5	54
25911_FFPE	3	200	2.58	195	99	61	2	7	53
40384_FFPE	3	200	1.61	234	114	56	2	5	54
37762_FFPE	3	200	0.04	232	131	58	2	5	53
32623_FFPE	3	200	2.2	174	64	51	2	13	51

FFPE: formalin-fixed paraffin embedded. ng: nanogram. [a] Reads that are not marked as duplicates. *CREv2 (Agilent). ΔΔCq: DNA quality score measured based on the FFPE NGS qPCR QC assay. [†] Bases that did not align the target region.

Table 5.8: Input DNA quality and quantity and whole-exome sequencing* data quality metrics of the germline samples from the ILBC methylation-defined subgroups calculated using *Picard*.

Germline sample (GC-DNA)	ILBC sub-group	Input DNA quantity (ng)	Input DNA quality (ΔΔCq)	Number of unique[a] reads (million)	Mean target* depth of coverage	Bases of target covered at least at 30X (%)	Bases of target not covered at all (%)	Bases off target[†] (%)	PCR duplication (%)
8965_GC	1	200	-0.7	70	49	54	2	7	28
34757_GC	1	200	-0.8	67	46	51	2	7	27
13412_GC	1	200	0.37	62	42	48	2	7	27
10780_GC	1	200	-0.63	70	49	54	2	7	28
27328_GC	1	200	-0.84	76	52	54	2	6	29
34778_GC	2	200	0.09	62	41	48	2	7	25
24634_GC	2	200	-0.01	59	38	43	2	7	25
29513_GC	2	200	-0.35	59	40	48	2	7	25
27217_GC	2	183	0.59	32	19	22	2	7	18
38184_GC	2	200	-0.15	41	24	28	2	6	23
33053_GC	3	200	0.16	56	36	44	2	7	23
25911_GC	3	200	-0.47	68	45	51	2	7	27
40384_GC	3	200	0.20	25	15	15	2	7	16
37762_GC	3	200	-0.39	35	22	26	2	6	21
32623_GC	3	200	-0.37	67	46	51	2	7	26

GC-DNA: DNA derived from dried blood spots stored on Guthrie Cards. ng: nanogram. [a] Reads that are not marked as duplicates. *CREv2 (Agilent). ΔΔCq: DNA quality score measured based on the NGS qPCR QC assay. [†] Bases that did not align the target region.

The input DNA quality, as measured by the ΔΔCq score, varied across the samples in the ILBC methylation-defined subgroups where a lower ΔΔCq score represents a sample with more amplifiable DNA. For germline samples, the ΔΔCq scores ranged from -0.84 to 0.59 (Table 5.8), whereas across the tumour samples, a wide variation was observed with the ΔΔCq scores ranging from 0.04 to 4.6 (Table 5.7). A significant negative correlation was observed between the ΔΔCq score and the mean target depth of coverage in both the tumour (Pearson correlation, R = -0.59, P-value = 0.02) and germline samples (R = -0.58, P-value = 0.04) (Figure 5.9b and Figure 5.10b). While, a significant negative correlation between the ΔΔCq score and the number of unique reads was observed in the germline DNA samples (R = -0.57, P-value = 0.04) (Figure 5.10a), the association was not statistically significant in the tumour DNA samples (R = -0.42, P-value = 0.12) (Figure 5.9a). The ΔΔCq score was negatively correlated with the percentage of target covered at a read depth of 30X in both the tumour and germline samples however, the association was not statistically significant for any of the sample types (Figure 5.9c, Figure 5.10c). While the ΔΔCq score showed a strong positive correlation with the PCR duplication rate (R = 0.74, P-value = 0.002) in the tumour, the association was negative in the case of germline samples (Figure 5.9e, Figure 5.10e).

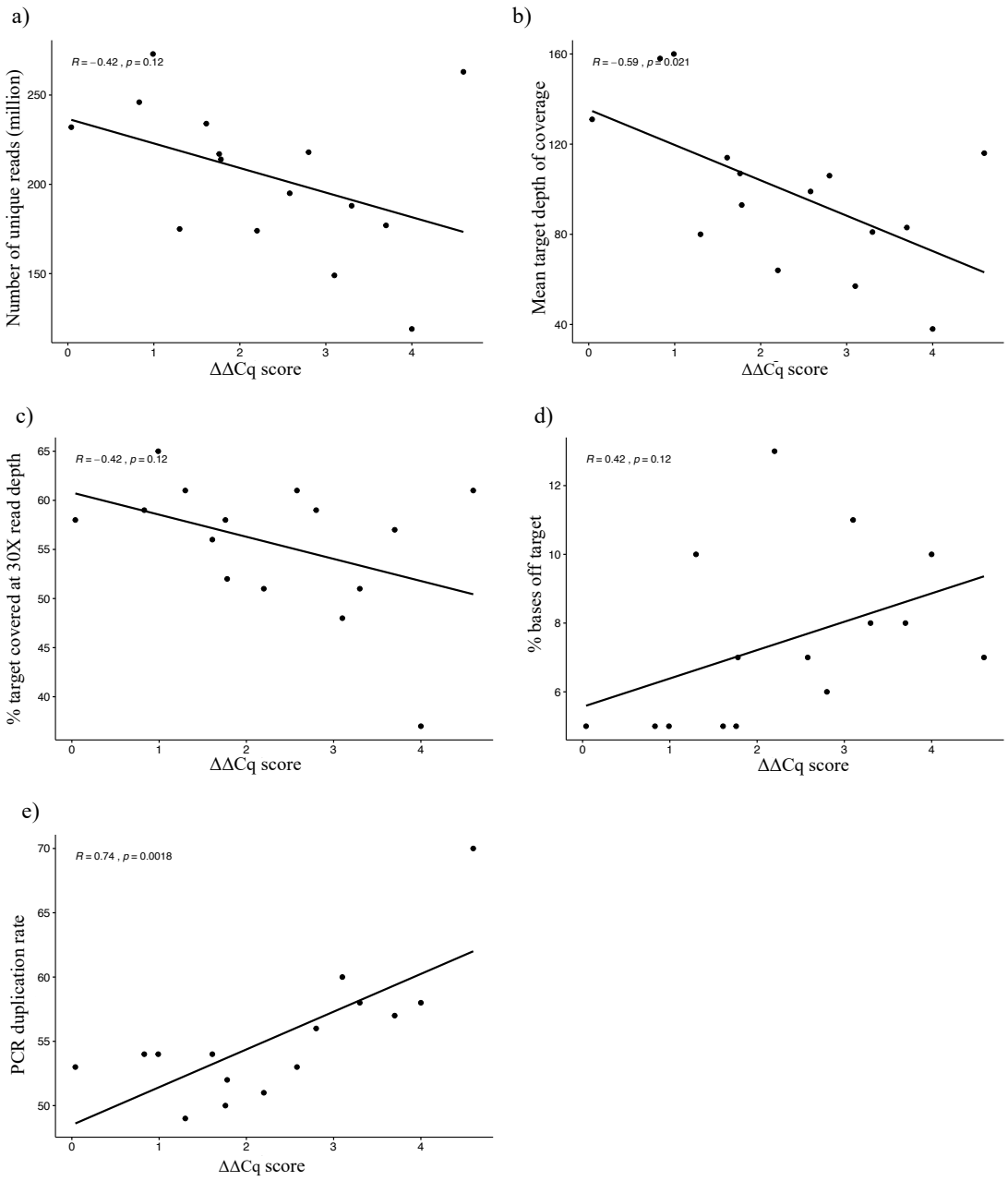


Figure 5.9: Correlation between the DNA integrity score measured as the ΔΔCq score and the sequencing data quality metrics of the tumour samples.

Scatterplots showing the correlation between the DNA quality measured as ΔΔCq score using the Agilent NGS qPCR QC assay (on the x-axis) and whole-exome sequencing metrics in the tumour samples (on the y-axis) where low ΔΔCq score represents a sample DNA with high amplifiability and vice-versa. The Pearson correlation coefficient, *R* and *P*-value indicating the strength and significance of correlation between the ΔΔCq score and **a**) Number of unique reads (million), *R* = -0.42, *P*-value = 0.12; **b**) Mean target depth of coverage, *R* = -0.59, *P*-value = 0.021; **c**) Percentage target covered at 30X read depth, *R* = -0.42, *P*-value = 0.12; **d**) Percent bases off-target, *R* = 0.42, *P*-value = 0.12 and **e**) PCR duplication rate, *R* = 0.74, *P*-value = 0.0018 are indicated in each plot in the top-left.

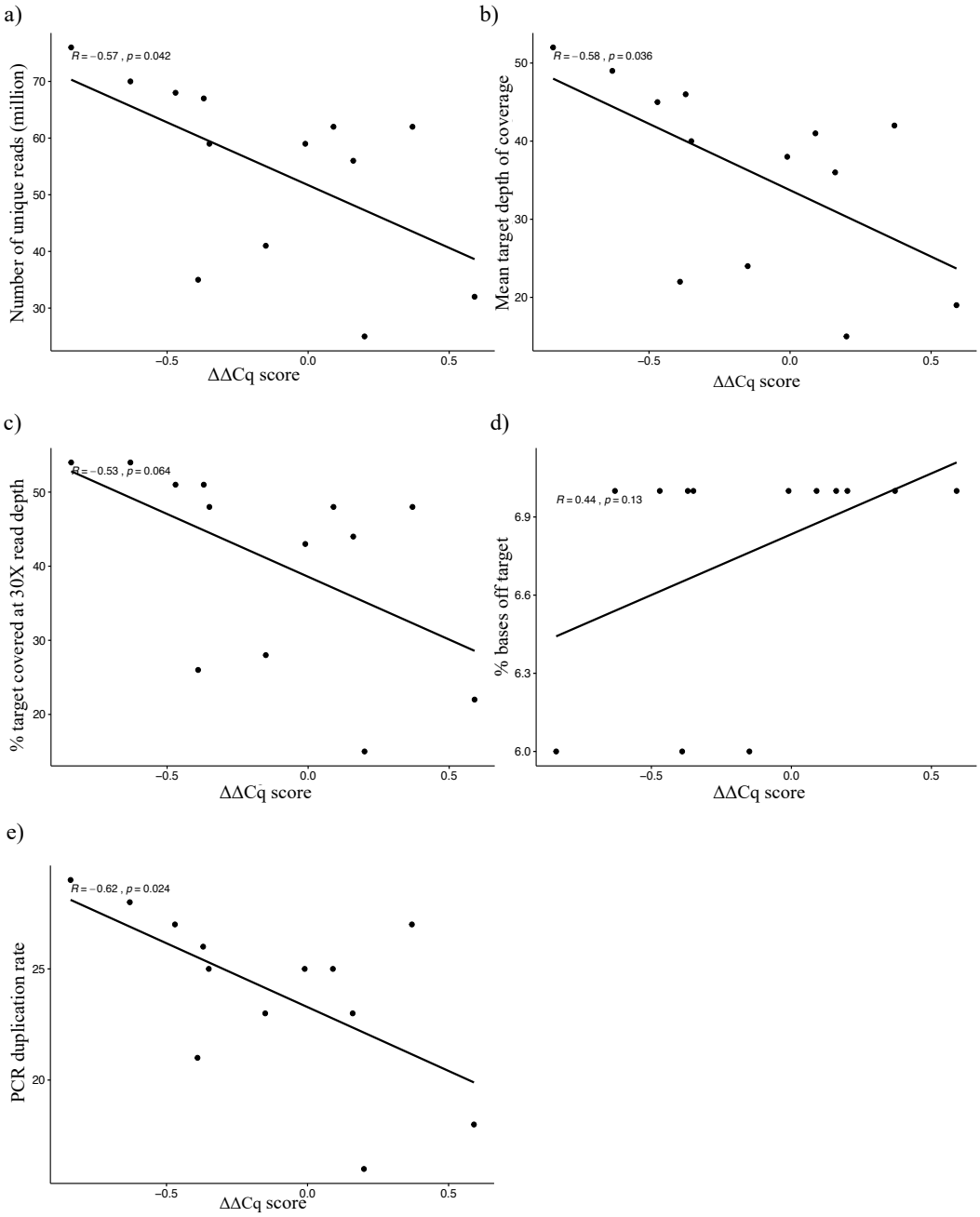


Figure 5.10: Correlation between the DNA integrity score measured as the ΔΔCq score and the sequencing data quality metrics of the germline samples.

Scatterplots showing the correlation between DNA quality measured as ΔΔCq score using Agilent NGS qPCR QC assay (on the x-axis) and sequencing data metrics of the germline samples (on the y-axis) where low ΔΔCq score represents a sample DNA with high amplifiability and vice-versa. The Pearson correlation coefficient, *R* and *P*-value indicating the strength and significance of correlation between the ΔΔCq score and **a**) Number of unique reads (million), *R* = -0.57, *P*-value = 0.042; **b**) Mean target depth of coverage, *R* = -0.58, *P*-value = 0.036; **c**) Percentage target covered at 30X read depth, *R* = -0.53, *P*-value = 0.064; **d**) Percent bases off-target, *R* = 0.44, *P*-value = 0.13 and **e**) PCR duplication rate, *R* = -0.62, *P*-value = 0.024 are indicated in each plot in the top-left.

The sequencing yielded 57 million unique reads for the germline samples and 205 million unique reads for the tumour samples on average. Mean target depths of 38X (15X-52X) and 99X (38X-160X) were achieved for the germline and tumour samples, respectively. On average 42% of the target was covered at a read depth of 30X in the germline samples, whereas 56% of the target was covered at a read depth of 30X in the tumour samples. A higher duplication rate ranging between 49%-70% (average 55%) was observed for the tumour samples when compared with the germline samples (16%-29%, average 25%). Deeper sequencing in the main experiment based on our findings in the pilot study did not seem to improve the target depth of coverage. The mean target depth of coverage for both the tumour and germline DNA was lower than the aimed target coverage, which were 150X mapped (44Gb of raw data/tumour sample) for the tumour and 40X-50X mapped (11Gb of raw data/per germline sample) for the germline samples.

5.6.2 Mutational signatures of tumours in ILBC methylation subgroups

A total of 17 different mutational signatures were identified in the ILBC tumours (n=15) from the ILBC methylation-defined subgroups (Figure 5.11), that included signatures associated with both exogeneous and endogeneous exposures (Table 5.9).

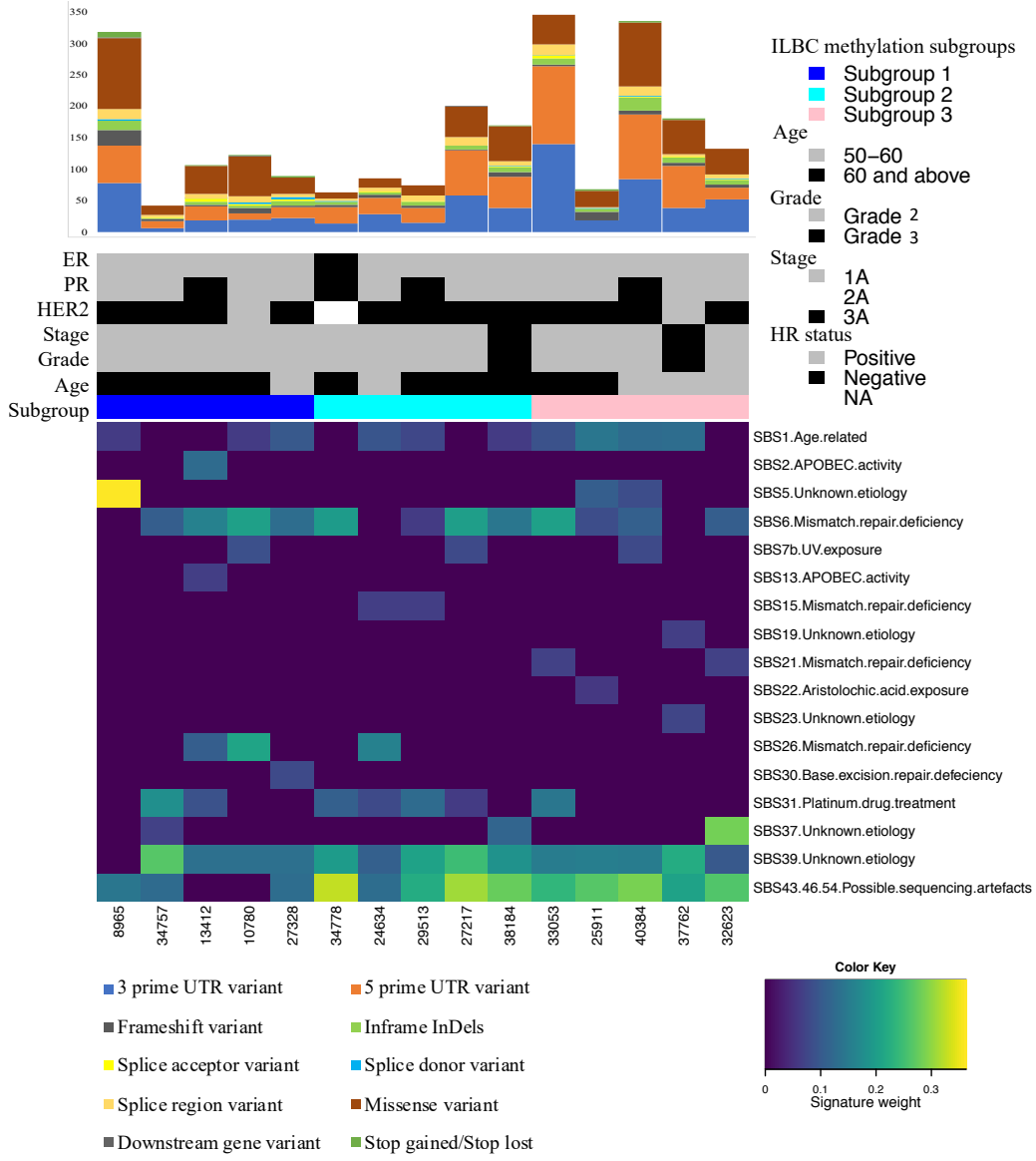


Figure 5.11: Mutational signatures identified in the tumours from the ILBC methylation-defined subgroups.

Heatmap showing the mutational signatures identified in the tumours of ILBC samples from the three ILBC methylation-defined subgroups, generated using *deconstructSigs* (Rosenthal *et al.*, 2016). The samples are plotted on the x-axis (as columns) and the mutational signatures are plotted on the y-axis (as rows). Signature weights are represented by colours in the heatmap as indicated in the colour key in the bottom-right corner. Stacked bar plot shows different types of somatic variants identified in the ILBC tumours as indicated in the legend in the bottom-left. The colour bar "Subgroup" indicates the three ILBC methylation-defined subgroups. The colour bars "Age", "Grade" and "Stage" indicate the age of women at tumour diagnosis, tumour grade and tumour stage, respectively. The colour bars "ER", "PR" and "HER2" indicate the expression status (measured by immunohistochemistry) of estrogen receptor, progesterone receptor and human epidermal growth factor receptor 2 of the ILBC tumours. The legends on the top-right indicate the different categories in the colour bars.

Signatures associated with exogeneous environmental exposures were SBS7b, associated with ultraviolet light-induced DNA damage (Pfeifer *et al.*, 2005), SBS22, associated with exposure to aristolochic acid (Poon *et al.*, 2013) and SBS31 associated with platinum drug treatment (Boot *et al.*, 2018) (Table 5.9). SBS6, associated with DNA mismatch repair deficiency (MMRd) was the most commonly observed endogeneous signature in the ILBC tumours, detected in 4/5 (80%) of the tumours in each ILBC subgroups. However, no significant difference in the mean signature weight of SBS6 was observed between the ILBC methylation-defined subgroups (ANOVA, *P*-value = 0.66). Other MMRd associated signatures were also displayed by the ILBC tumours that included SBS15, observed in 2/5 (40%) of the tumours in Subgroup 2, SBS21, observed in 2/5 (40%) of the tumours in Subgroup 3 and SBS26 observed in 2/5 (40%) of the tumours in Subgroup 1 (Table 5.9). No significant difference in the cumulative weight of MMRd associated mutational signatures SBS6, SBS15, SBS21 and SBS26 was observed between the three subgroups (ANOVA, *P*-value = 0.24). Although, MMRd associated signatures were observed ubiquitously in ILBC tumours across all three subgroups, no somatic or pathogenic germline variants in the mismatch repair genes *MLH1*, *MSH2*, *MSH3*, *MSH6*, *PMS1* and *PMS2* were observed in any of these cases according to ClinVar (accessed on 2020-01-01).

SBS2 and SBS13, both associated with apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like (APOBEC) cytidine deaminase DNA-editing activity endogeneous exposures (Nik-Zainal *et al.*, 2012), were observed in only one tumour from Subgroup 1 (Table 5.9). SBS30, associated with base excision repair deficiency was observed in 1/5 (20%) of the tumours in Subgroup 1. Several signatures were identified in the ILBC tumours for which the etiology is currently unknown that included SBS5,

SBS19, SBS23, SBS37 and SBS39. Of these, SBS39 was the most prominent and was detected in 14/15 (93%) of the ILBC tumours (Table 5.9).

Table 5.9: Mutational signatures identified in the tumours of the ILBC methylation-defined subgroups.

	Mutational signature	Subgroup 1 (n=5)	Subgroup 2 (n=5)	Subgroup 3 (n=5)
		Number of cases, n (%)		
Exogeneous exposure	SBS7b-UV exposure	1 (20)	1 (20)	1 (20)
	SBS22-Aristolochic acid exposure	0 (0)	0 (0)	1 (20)
	SBS31-Platinum drug treatment	2 (40)	4 (80)	1 (20)
Endogeneous exposure	SBS1-Age related	3 (60)	3 (60)	4 (80)
	SBS2 and SBS13-APOBEC activity	1 (20)	0 (0)	0 (0)
	SBS6-Mismatch repair deficiency	4 (80)	4 (80)	4 (80)
	SBS15-Mismatch repair deficiency	0 (0)	2 (40)	0 (0)
	SBS21-Mismatch repair deficiency	0 (0)	0 (0)	2 (40)
	SBS26-Mismatch repair deficiency	2 (40)	0 (0)	0 (0)
	SBS30-Base excision repair deficiency	1 (20)	0 (0)	0 (0)
Unknown etiology	SBS5	1 (20)	0 (0)	2 (40)
	SBS19	0 (0)	0 (0)	1 (20)
	SBS23	0 (0)	0 (0)	1 (20)
	SBS37	1 (20)	1 (20)	1 (20)
	SBS39	4 (80)	5 (100)	5 (100)

5.6.3 Mismatch repair deficiency and microsatellite instability

The mismatch repair system plays a key role in repairing the mismatched nucleotides and thus, maintain genomic stability and integrity and prevents insertion and deletions of DNA at the microsatellite sites (Hsieh & Yamane, 2008). Due to its role in genomic stability, some of the downstream consequences of MMRd are microsatellite instability, high mutation load and the accumulation of large number of InDel variants in tumours (Hsieh & Yamane, 2008).

The MSI-score of the ILBC tumours ranged from 0 to 58 (Table 5.10). Based on the MSI-score, 13/15 (87%) of the ILBC tumours were classified as microsatellite instable. Comparing the correlation between the cumulative weight of MMRd associated signatures SBS6, SBS15, SBS21 and SBS26, and the MSI-score calculated using *MSIsensor,* no significant correlation was observed between the two (Figure 5.12).

Table 5.10: Percentage of the target region covered at a minimum read depth of 30X and the effective target region used for calculating the tumour mutation burden and summary of different measures estimated in the tumour from the ILBC methylation-defined subgroups.

Tumour sample	ILBC methylation-defined Subgroup	Bases of target* covered at least at 30X read depth (%)	Effective target region of total 67.3Mb* for TMB[a] calculation (Mb)	TMB[a]	InDel variants per Mb of target region covered at 30X read depth	Cumulative weight of MMRd associated mutational signatures	MSI-score[b]
8965	1	48	32	46	12.2	0.08	7
34757	1	59	40	5	0.9	0.19	0
13412	1	61	41	11	1.6	0.26	26
10780	1	37	25	19	3.4	0.42	18
27328	1	51	34	10	1.6	0.15	58
34778	2	59	40	9	1.9	0.2	7
24634	2	61	41	11	1.9	0.26	0
29513	2	57	38	7	1.8	0.17	9
27217	2	58	39	24	4.2	0.2	22
38184	2	52	35	22	3.1	0.15	14
33053	3	65	44	57	9.4	0.28	13
25911	3	61	41	9	1.5	0.11	28
40384	3	56	38	42	6.2	0.11	21
37762	3	58	39	22	3.9	0.08	17
32623	3	51	34	42	2.3	0.18	33

ILBC: Invasive lobular breast cancer. * Clinical research exome v2 (Agilent), 67.3Mb. [a] Tumour mutation burden. InDel: Insertion and deletion. MMRd: Mismatch repair deficiency. [b] Microsatellite instability score calculated using *MSI-sensor*.

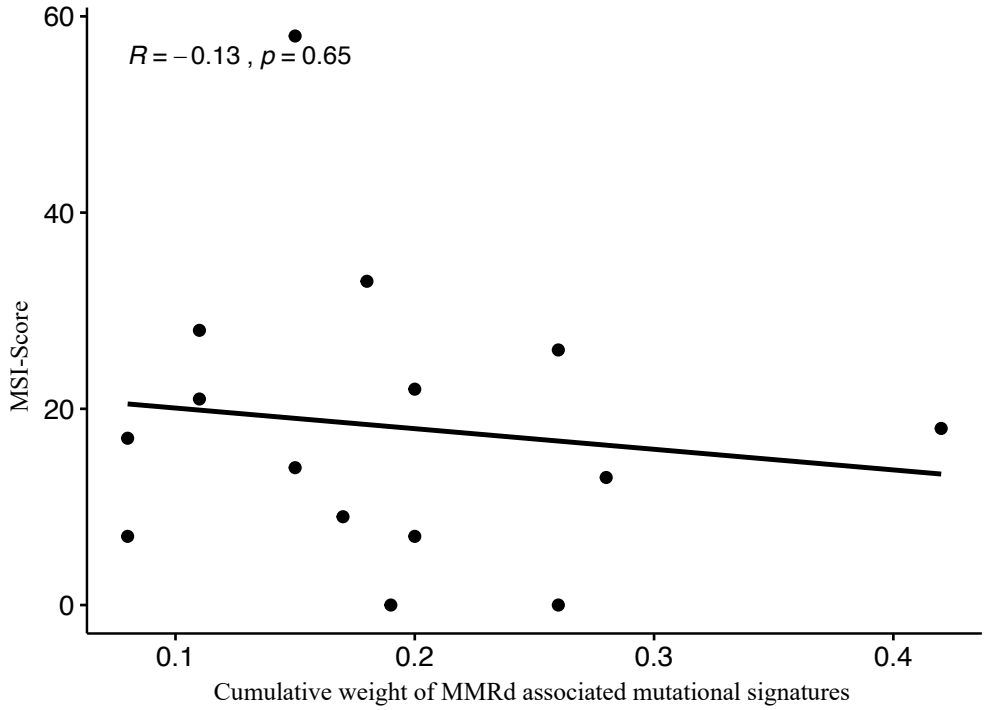


Figure 5.12: Correlation between the cumulative weight of mismatch repair deficiency associated mutational signatures and the microsatellite instability score of the tumours from the ILBC methylation-defined subgroups.

Scatterplot showing the correlation between the cumulative weight of mismatch repair deficiency (MMRd) associated mutational signatures SBS6, SBS15, SBS21 and SBS26 (on the x-axis) and the microsatellite instability score (MSI-score) calculated using *MSIsensor* (Niu *et al.*, 2014) (on the y-axis) of the tumours in the ILBC methylation-defined subgroups. The correlation coefficient, R = - 0.13 and P-value = 0.65, indicate the strength and significance of the association and are also marked on the top-left of the plot.

The TMB of the ILBC tumours was calculated after normalising to the actual target region covered by each sample at a minimum read depth of 30X (the applied cut-off for variant filtering) (Table 5.10). The TMB of ILBC tumours in the methylation-defined subgroups ranged from 5 to 56 mutations/Mb (Table 5.10). The average TMB in Subgroup 1 was 17 mutations/Mb compared with 14 mutations/Mb in Subgroup 2. Tumours in Subgroup 3 had the highest TMB ranging from 8 to 56 (average 33 mutations/Mb) however, there was no significant difference in the mean TMB between

the three subgroups (ANOVA, *P*-value = 0.13). The rate of InDel variants (total number of InDel variants detected per Mb of target region covered at a minimum read depth of 30X) was also higher in Subgroup 3 (mean=4.7) when compared with Subgroup 1 (mean=3.9) and Subgroup 2 (mean=2.6) (Table 5.10). While no significant correlation was observed between the cumulative weight of the MMRd associated mutational signatures and the TMB of the tumours (Figure 5.13a), the TMB showed a strong positive correlation with the rate of InDel variants (R = 0.86, P-value = 4.3×10^{-5}) (Figure 5.13b).

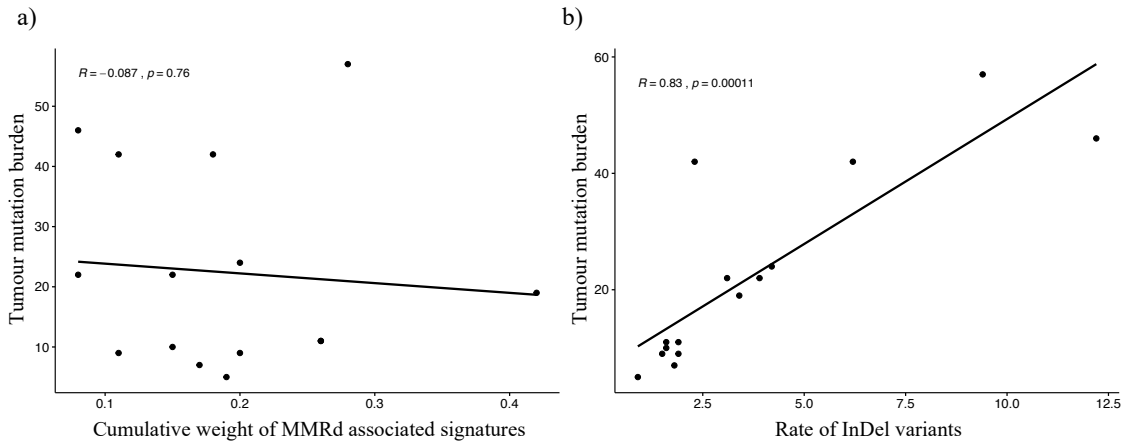


Figure 5.13: Correlation between the tumour mutation burden of tumours from the ILBC methylation-defined subgroups with mismatch repair deficiency associated mutational signatures and total number of somatic insertion and deletion variants in the tumours.

Scatterplots showing the correlation between **a**) the cumulative weight of mismatch repair deficiency (MMRd) associated signatures (on the x-axis) and the tumour mutation burden (TMB) (on the y-axis), R = -0.087, P-value = 0.76 and **b**) the TMB and the rate of somatic insertion and deletion (InDel) variants (R = 0.83, P-value = 0.00011), identified in the tumours from the ILBC methylation-defined subgroups. The correlation coefficient, R and P-value indicating the significance of the association are also marked on the top-left of the plot.

5.6.4 Methylation status of mismatch repair deficiency related genes

As promoter hypermethylation of mismatch repair genes, in particular *MLH1*, has been reported to be associated with MMRd (Bevilacqua & Simpson, 2000; Kuismanen *et al.*, 2000; Salvesen *et al.*, 2000), the association between the methylation status of mismatch repair genes *MLH1*, *MSH2*, *MSH3*, *MSH6*, *PMS1* and *PMS2*, and MMRd associated mutational signatures were investigated. The ILBC tumours in all three methylation-defined subgroups showed a consistent hypomethylation pattern (beta-value < 0.50) across the promoter regions of *MSH2*, *MSH3*, *MSH6*, *PMS1* and *PMS2* (Figure 5.15). However, the methylation pattern across the TSS1500 region of *MLH1* (14 CpGs, 973 bp) was variable across the ILBC tumours (Figure 5.14). The variable region overlapped the functional promoter region of *MLH1* as described in (James G Herman *et al.*, 1998) (GenBank U83845). Hypermethylation (beta-value > 0.50) was observed at six CpG positions across this variable region with 3 CpGs; cg02103401, cg24607398 and cg10990993 found to be hypermethylated in 9/15 (60%), 11/15 (73%) and 12/15 (80%) of the ILBC tumours. However, DNA methylation level at these 14 CpG positions located in the TSS1500 region did not show any significant correlation with the cumulative weight of MMRd signatures weight and MSI-score of the ILBC tumours (Table 5.11).

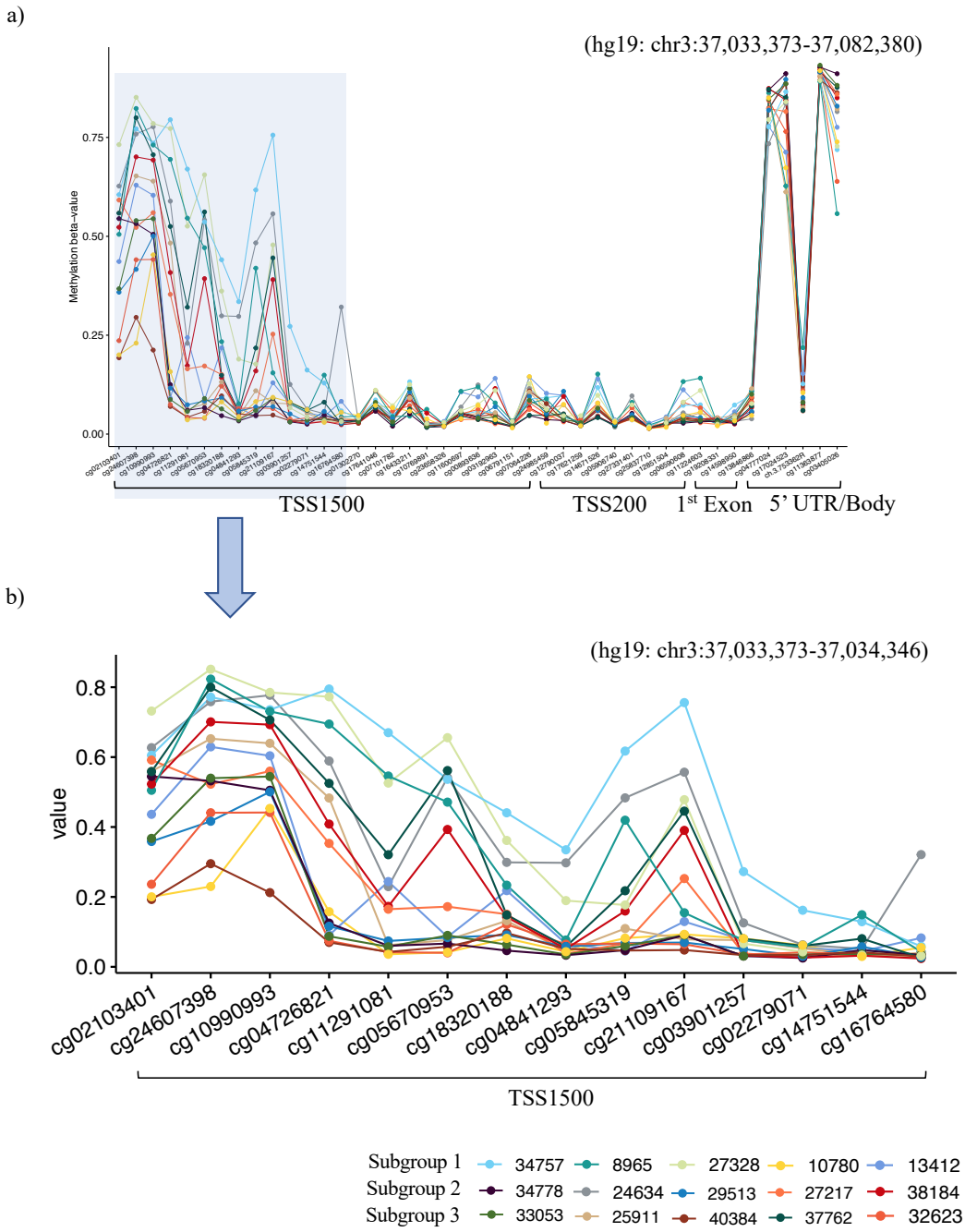


Figure 5.14: Methylation patterns of ILBC tumours at *MLH1*.

Plots showing the methylation patterns of tumours from the ILBC methylation-defined subgroups at the promoter region of *MLH1* **a**) shows the total *MLH1* gene region covered by the CpG probes in the assay and **b**) shows a higher resolution of the TSS1500 region of the *MLH1* gene. The CpG positions overlapping different genomic locations of the gene are shown on the x-axis and the methylation level (beta-value) of the samples are shown on the y-axis. Genomic coordinates for the regions are marked in the corresponding plot in the top-left corner. The different colour lines represent different samples belonging to the ILBC methylation-defined subgroups as indicated in the legend on the bottom right.

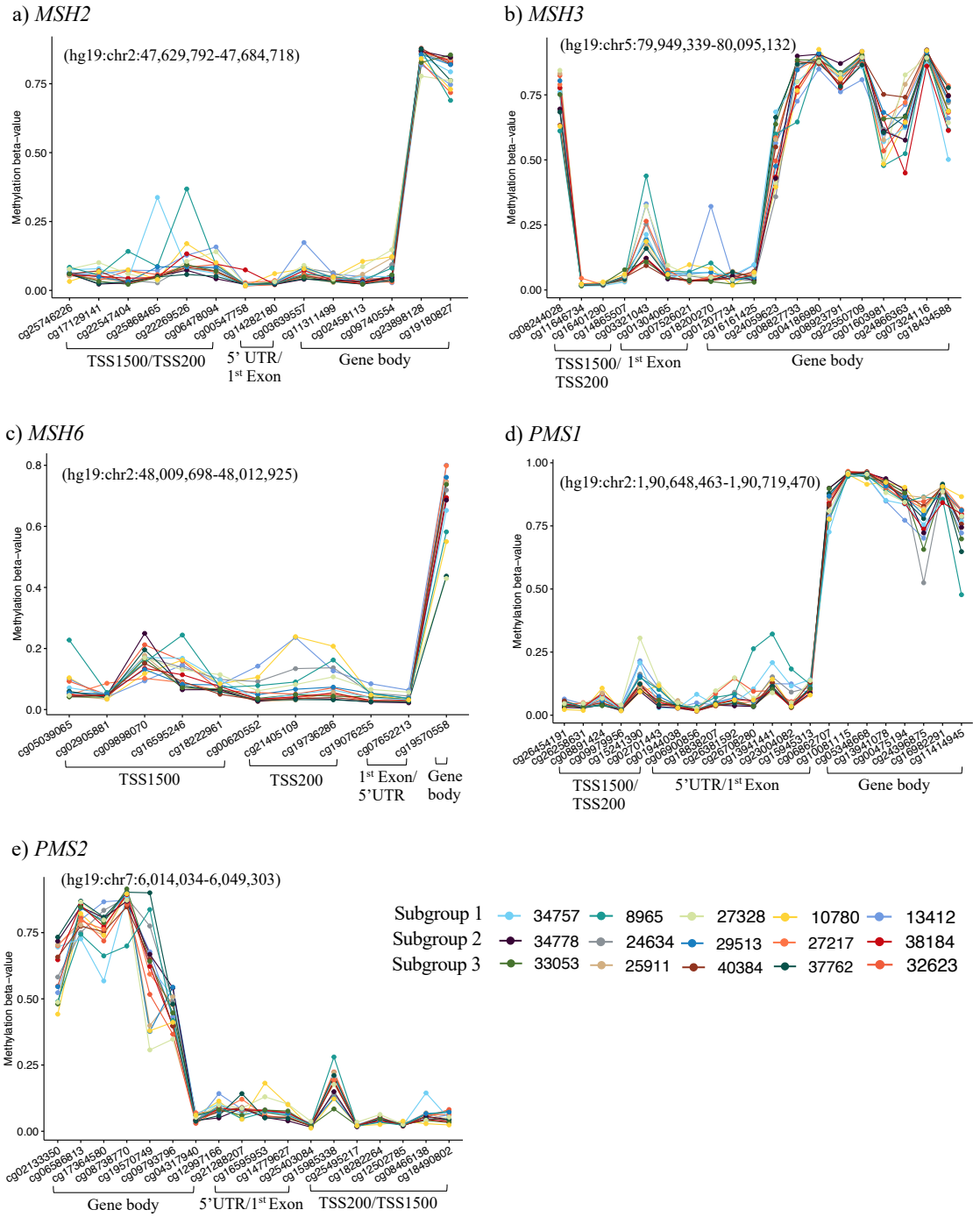


Figure 5.15: Methylation pattern of ILBC tumours at mismatch repair genes.

Plots showing the methylation patterns of ILBC tumours from the methylation-defined subgroups across the genes associated with mismatch repair; **a**) *MSH2* **b**) *MSH3* **c**) *MSH6* **d**) *PMS1* and **e**) *PMS2*. Similar data for *MLH1* is shown in the **Figure 5.14**. The CpG positions overlapping different genomic locations of the genes are shown on the x-axis and the methylation level (beta-value) of the samples are shown on the y-axis. Genomic coordinates for the regions are marked in the corresponding plot in the top-left corner. The different colour lines represent different samples belonging to the ILBC methylation-defined subgroups as indicated in the legend on the bottom right.

Table 5.11: Correlation between DNA methylation at each CpG position in the TSS1500 region of *MLH1* and mismatch repair deficiency and microsatellite instability score of the ILBC tumours.

		MMRd[a]		MSI-Score[b]	
CpG position (TSS1500* of *MLH1*)	**Number of hypermethylated ILBC tumours, n (%)**	***R***	***P*-value**	***R***	***P*-value**
cg02103401	9 (60)	-0.33	0.23	0.059	0.84
cg24607398	11 (73)	-0.47	0.076	0.023	0.94
cg10990993	12 (80)	-0.17	0.54	-0.037	0.9
cg04726821	5 (33)	-0.38	0.16	0.01	0.97
cg11291081	3 (20)	-0.32	0.25	-0.011	0.97
cg05670953	4 (27)	-0.35	0.21	-0.006	0.98
cg18320188	0 (0)	-0.13	0.64	0.062	0.83
cg04841293	0 (0)	0.052	0.85	-0.2	0.47
cg05845319	1 (7)	-0.15	0.6	-0.46	0.083
cg21109167	1 (7)	-0.11	0.71	-0.15	0.58
cg03901257	0 (0)	0.074	0.79	-0.36	0.18
cg02279071	0 (0)	0.075	0.79	-0.32	0.24
cg14751544	0 (0)	-0.38	0.17	-0.49	0.065
cg16764580	0 (0)	0.31	0.26	-0.35	0.2

* Genomic region from -200 to -1500 nucleotides upstream of the transcription start site. ILBC: Invasive lobular breast cancer. R: Pearson's correlation coefficient. [a] Mismatch repair deficiency. [b] Microsatellite instability score calculated using *MSIsensor*.

5.7 Discussion

This chapter presents the results of the investigation of the exomic somatic mutation profiles of ILBC tumours within a selection of cases from the DNA methylation defined ILBC subgroups identified in Chapter 4. Mutational signatures present in the ILBC tumours suggested a potential involvement of MMRd in the tumourigenesis and progression of some of these ILBCs. Since WES using FFPE DNA was not previously used in our lab, a pilot study was conducted. Data from the pilot study was evaluated to test the effect of DNA quality and quantity on the sequencing data quality and the suitability of GC-DNA to be used as the germline reference sample for detecting somatic variants and identifying mutational signatures.

Input DNA quality and quantity are important factors to consider while preparing libraries for next generation sequencing experiments. Comparison of the sequencing metrics of two tumour DNA samples with different quantities of input DNA from the pilot study showed that Sample 2-FFPE-rep with higher input DNA, showed a nominally higher depth and breadth of coverage compared with Sample 2-FFPE (Table 5.3). A higher input DNA quantity also seemed to reduce the PCR duplication rate. PCR duplicates represent the fraction of sequencing reads that arise from the same original DNA template. One of the main reasons for high PCR duplication rates is low starting input DNA amount that leads to over-amplification of the same DNA template during library preparation (Akbari *et al.*, 2005). This is reflected in the case of Sample 2-FFPE and Sample 2-FFPE-rep. Although, a better sequencing metrics was observed with increased amount of input DNA, the difference was not large. However, this may be because this comparison involved two aliquots of the same DNA extraction. Two DNA samples with the same DNA quality may have minimised the impact of different DNA

quantities. Although the library preparation kit by Agilent is designed for low input DNA, the pilot data suggested an improvement in the sequencing data quality when higher amount of input DNA was used for library preparation. In the main experiment, the DNA quality (ΔΔCq score) showed a large variation, ranging from 0.04 to 4.6 for the tumour DNA samples (Table 5.7) and from -0.84 to 0.59 for the germline DNA samples (Table 5.8). Therefore, based on the evaluation of Sample 2-FFPE and Sample 2-FFPE-rep, 200 ng of input DNA was used in the main experiment for all the samples. However, it should be noted that this conclusion was based on the comparison of only two samples.

DNA extracted from FFPE tumour tissues are highly degraded and fragmented. Formalin fixation reduces the efficiency of FFPE derived DNA for PCR amplification due to DNA-protein crosslinks that increases the sensitivity of DNA to mechanical stress and decreases its accessibility for enzymes (Do & Dobrovic, 2015). Excessive fragmentation could also lead to FFPE-DNA fragments that are too small for Illumina bridge amplification (Bentley *et al.*, 2008). Therefore, an accurate quality and quantity estimation of amplifiable DNA is crucial before library preparation for obtaining uniform data quality and coverage (Quach *et al.*, 2004; Do & Dobrovic, 2015). The Qubit assay quantifies the absolute amount of double stranded DNA present in a sample that also includes DNA fragments that are not suitable for amplification. The FFPE NGS qPCR QC assay (Agilent), which was used in this study for DNA quantification, measures the concentration of amplifiable DNA fragments. It is a PCR-based assay that amplifies two different genomic regions and generates amplicons of sizes, 42 bp and 123 bp. The DNA integrity score (ΔΔCq) is calculated by comparing the cycle threshold values (Cq) for amplicon A (42bp) and amplicon B (123bp) of the sample with a reference DNA with high molecular weight. A lower ΔΔCq score represents a sample with more amplifiable

DNA. The input DNA quantity was adjusted based on the ΔΔCq score before library preparation as per manufacturer's instructions. For samples with ΔΔCq score > 1, representing sample DNA with low amplifiability, the input DNA quantity was calculated based on the qPCR results thus, considering the exact amount of amplifiable DNA in the calculation. In addition to being a measure for DNA quality and amplifiability, ΔΔCq score was also found to be predictive of the sequencing data quality in the tumour and germline DNA samples from the ILBC methylation-defined subgroups. High ΔΔCq score representing samples with lower amplifiability were found to be associated with lower mean target depth of coverage and higher PCR duplication rates in the tumour DNA samples. Therefore, the ΔΔCq score represents an important predictor of sequencing performance at the very initial stage of the assay and may be used to make informed decisions regarding sample selection.

To evaluate the robustness and sensitivity of the WES assay, the sequencing data qualities of the tumour samples and their technical replicates from the pilot study were compared. We did not observe a large difference in the sequencing metrics of Sample 1-FFPE and Sample 1-FFPE-rep (Table 5.3). Small variation observed in the target depth of coverage and percent of target covered at a read depth of 30X may be related to differences in sample handing during library preparation. Variations introduced due to differences in sample handling during library preparation is evident in the case of Sample 1-FFPE and Sample 1-FFPE-rep (Figure 5.5). Looking at their post-hybridisation library profiles (Figure 5.5a, Figure 5.5b), differences in the library fragment size distribution can be noted between the profiles of the same tumour DNA sample. Cluster generation and sequencing efficiency on Illumina systems are known to be influenced by the library fragment size (Head *et al.*, 2014). Smaller fragments tend to cluster more efficiently to

the flow cell than the larger fragments. This may partly explain small differences observed in the mean target depth of coverage and percentage of target covered for Sample 1-FFPE and Sample 1-FFPE-rep. In the case of Sample 2-FFPE and Sample 2-FFPE-rep, a slightly better sequencing performance was observed for Sample 2-FFPE-rep that may be attributed to the higher amount of input DNA used for this sample. As discussed earlier, higher input DNA was not found to have a large impact on the sequencing data quality as this comparison involved two aliquots of the same DNA sample thus, two samples with the same DNA quality.

Since, the tumour DNA samples and their respective technical replicates were the duplicate aliquots of the same stock sample, we expected to see a good concordance between the SSNVs identified in them. However, comparing the SSNVs between the tumour samples and their respective replicate samples showed a low overlap. Between Sample 1-FFPE and Sample 1-FFPE-rep, only 41% of $SSNVs^{WB}$ and 53% of the $SSNVs^{GC}$ were concordant (Figure 5.7a, Figure 5.7b). In the case of Sample 2-FFPE and Sample 2-FFPE-rep**,** only 64% of $SSNVs^{WB}$ and 58% of $SSNVs^{GC}$ were concordant (Figure 5.7c, Figure 5.7d). One of the possible explanations of this high discordant rate observed between the tumour samples and their respective technical replicates could be the low target overlap observed between the samples. At a minimum read depth of 30X (the applied threshold for variant filtering), only 59% of the target bases overlapped between both Sample 1-FFPE and Sample 1-FFPE-rep and 65% of the target bases overlapped between both Sample 2-FFPE and Sample 2-FFPE-rep. The SSNVs detected only in the common target region were compared to calculate the concordance rate between the samples thus, resulting in a high observed discordance. Other reason for the discordance could be the sequencing artefacts that may be present even after variant

filtering. Since the tumour DNA showed a range of DNA quality (ΔΔCq score) and the percentage of target covered at a minimum read depth of 30X also varied across the tumour samples, the applied variant filters were not suitable for all the samples. The concordance level between the tumour DNA samples and their technical replicates was low (~3%) before any variant filtering and although, a considerable improvement in the concordance rates between the tumour samples and their technical replicates was observed after variant filtering (minimum read depth of 30X and minimum variant allele fraction of 0.2), artefacts still remained. Applying a higher read depth and allele frequency filters may have further reduced the sequencing artefact rates however, this would also lead to the loss of a large portion of target region as the percentage of target covered reduced with increasing read depths for both the tumour and the germline reference samples (44% target covered at 50X read depth and 21% at 100X read depth for FFPE *versus* 24% at 50X and 8% at 100X read depth for germline, GC-DNA and WB-DNA samples). Applying higher thresholds for variant filtering would also increase the chance of removing some true variants present at lower variant allele fractions and depth and hence more stringent filtering cut-offs were not applied. Poor sequence read depth in the germline reference sample may lead to calling germline events as somatic. To mitigate this, variant filtering may be performed based on whether the variants were reported in dbSNP to remove germline variants and to get a more refined list of somatic variants. There are other approaches that could be applied to further refine the somatic variant identification such as variant calling using two different somatic variant caller and using the common variants identified by the two callers as the final set for downstream analyses.

The suitability of dried blood spots stored on Guthrie Cards to be used as the matching germline reference sample was tested in the pilot study. Guthrie Cards provide

a means for long term blood storage that has been proved to be an efficient and cost-effective alternative to conventional blood freezing for genomic and epigenomic studies (Mei *et al.*, 2001; H He *et al.*, 2007; Mas *et al.*, 2007; Al Safar *et al.*, 2011; Ghantous *et al.*, 2014; Nguyen-Dumont *et al.*, 2015). Blood samples collected on a GC can be stored at room temperature indefinitely prior to DNA extraction, thus reducing the cost and infrastructure requirements for low temperature storage. The overall sequencing performance of GC-DNA and WB-DNA was similar for both the ILBC cases in the pilot study (Table 5.3). However, comparing $SSNVs^{WB}$ and $SSNVs^{GC}$, a low concordance level of approximately 40% and 55% was observed between the two variant call sets for Sample 1-FFPE and Sample 2-FFPE, respectively (Figure 5.3). The low concordance level between $SSNVs^{WB}$ and $SSNVs^{GC}$ could be explained by the difference in the target region commonly covered by the WB-DNA and GC-DNA. At a minimum read depth of 30X, only 33% of the target region was found to be commonly covered by both Sample 1-WB and Sample 1-GC and only 34% of the target region was commonly covered by both Sample 2-WB and Sample 2-GC (Table 5.4). Since the overlapping target region between the two sources of germline reference samples was low (over 30% for both the samples), it resulted in the identification of different sets of somatic variants with low concordance, when the two sources of germline DNA were used as the reference samples. Although, considerable discordance was observed between $SSNVs^{WB}$ and $SSNVs^{GC}$, the expected mutational signature, SBS3 was detected as a primary signature in $SSNVs^{GC}$ for both Sample 2-FFPE and Sample 2-FFPE-rep that supported the suitability of GC-DNA to be used as the germline reference sample in somatic variant identification and mutational signature analysis.

The evaluation of the mutational signatures identified in the tumours from the ILBC methylation-defined subgroups, suggested a common occurrence of MMRd associated signatures SBS6, SBS15, SBS21 and SBS26 in the ILBC tumours. SBS6 was displayed by 12/15 (80%) of the ILBC tumours suggesting a possible mismatch repair defects in these tumours. However, it could not be supported by the occurrence of any somatic or germline pathogenic variant in the mismatch repair genes in any of these samples. This could, however, be due to the limited sensitivity of the assay as genetic variations (both somatic and germline) of variant allele fraction of 0.2 or more were confidently detected in this study and variants present at lower variant allele fraction within the mismatch repair genes could not be sensitively detected. Although the presence of MMRd associated signatures was not explained by any genetic variants in the mismatch repair genes in this study, we found hypermethylation (average beta-value > 50%) at three CpG positions (259 bp) located in the TSS1500 of *MLH1* in more than 50% of the ILBC tumours. The functional promoter of *MLH1* has been well-characterised and described to be located proximal to the TSS of *MLH1*. Hypermethylation at this region was found to be associated with loss of *MLH1* expression (James G Herman *et al.*, 1998; Deng *et al.*, 1999). Although DNA methylation at the CpG positions located in the TSS1500 region of *MLH1* did not show any correlation with the MMRd somatic profile and MSI status of the tumours, it does not invalidate the impact of hypermethylation at these CpGs on the gene expression of *MLH1*. Microsatellite instability was indicated in 13/15 (87%) of the ILBC tumours. Features of MMRd and microsatellite instability have been previously reported in breast cancers (Yee *et al.*, 1994; Shaw *et al.*, 1996; Murata *et al.*, 2005; Kappil *et al.*, 2016; Malik *et al.*, 2019). Murata et. al., (2002), identified somatic genetic variants in *MSH2* and hypermethylation at *MLH1* promoter as the two main alterations contributing to MMRd in breast cancers (Murata *et al.*, 2002). In another study, Murata

et al., (2005), reported a reduced expression of *MLH1* and *MSH2* in 31% (26/83) and 28% (23/83) of the sporadic breast cancer cases, respectively, with loss of *MLH1* expression predominantly caused by promoter hypermethylation in these cases. Microsatellite instability has been previously reported in ILBCs by two different studies (Aldaz *et al.*, 1995; Contegiacomo *et al.*, 1995). Aldaz et al., (1995) reported a significantly higher frequency of MSI in ILBC cases (39%, 9/23) compared with ductal breast cancers (14%, 7/52), (*P*-value= 0.012) (Aldaz *et al.*, 1995). Another group reported an association between MSI with lobular histology and increased lymph node involvement (Contegiacomo *et al.*, 1995). Both these studies used MSI markers to determine the microsatellite instability status of the tumour as opposed to computational approach used in this study. The mutational signature analysis of the ILBC samples suggested a possible role for MMRd in ILBC tumourigenesis. However, the data did not allow us to confirm the finding by associating the signatures with genetic variations in genes involved in MMR. Since many of the mutational signatures detected were driven predominantly by C>T substitution, which is also associated with formalin fixation, it may be possible that FFPE-induced artefact confounded the accurate signature assignment. However, the presence of expected signature SBS3 in Sample 2 from the pilot study gave us some reassurance that the mutational signatures algorithm was able to detect true signatures despite of the noise present in the data. Despite its limitations this study presents as an interesting base for future studies. Testing the ILBC tumours for MMRd using immunohistochemistry could be the next step forward. As artefactual variants bias the mutational profiles, it is crucial to introduce methodologies to reduce FFPE-induced artefacts in future sequencing for accurate unravelling of mutational signatures.

One of the main limitations of this study was low mean target depth of coverage and a low percentage of target coverage, which considerably reduced the breadth of analysis that could be done using the data. On an average, somatic variants were confidently detected only across 55% of the target region, which greatly limited the analysis to a small genomic region. The highly fragmented FFPE DNA samples further affected the data quality by introducing a large number of sequencing artefacts that were present at a variant allele fraction > 20%, as demonstrated by the sequencing artefacts associated mutational signatures in the tumour samples. Although we made adjustments in the sequencing data yield and additional sequencing was added to account for PCR duplication, overlapping reads and off-target coverage based on the results from the pilot study however, it was not reflected in the sequencing output of the main experiment. The mean target depth of coverage was lower (mean 99X for tumour DNA and 38X for germline DNA sample) than expected (~150X mapped for tumour DNA and 40X-50X mapped for germline DNA sample). One reason for poor yield may be the poorer FFPE DNA quality of the samples in the main experiment compared with the samples in the pilot study. Since the complexity of the library was poor, i.e., the proportion of unique DNA templates in the library was low, increasing the sequencing data did not improve the depth of target coverage. This could suggest that Agilent SureSelect XT low input kit may not be the best suited kit for poor quality FFPE tumour DNA sample type. Library preparation kits that use PCR free library preparation workflow could be a better suited chemistry for maximising the data quality obtained from the FFPE samples. However, removing PCR amplification may increase the input DNA requirement, which is often limited in the case of clinical FFPE samples. Some of the parameters that could also affect sequencing data quality and therefore somatic variant calling are DNA insert size, mapping quality and percentage of overlapping reads. Both the pilot study and the main

experiment generated 150 bp paired end reads, which may not be the most appropriate sequencing read length for the FFPE DNA as they are highly fragmented. Since the FFPE DNA fragments size are smaller that results in smaller insert size, selecting a sequencing read length of 150 bp leads to a higher percentage of overlapping reads. As target coverage was non-uniform across the samples, low concordance in somatic variants were observed between the FFPE samples and their replicates. Across the overlapping target region, the FFPE samples and their replicates had many unique variants that were recorded only in one of the replicates possibly because of sufficient depth in one replicate but not in another. Some of these limitations could be overcome by using DNA repair kits for repairing FFPE-derived DNA before sequencing. Further detail of the quality metrics discussed above are provided in appendix (Additional Table 1).

5.8 Summary

Although this study was limited by the sequencing data quality, a possible association between MMRd and ILBC development and progression has been indicated that may be related to *MLH1* promoter hypermethylation. Mismatch repair genes have been studied extensively in colorectal cancers however, the involvement of MMRd in ILBC is not well-characterised. Experimental validation of the MSI findings presented here will be an important future work. Considering the potential for immunotherapy in the context of MSI high tumours, it is possible that at least a subset of ILBC could benefit from these therapeutic approaches. Further research could look at the mismatch repair genes in larger studies for testing the germline DNA to determine if the mismatch repair genes should be on panel tests for ILBC susceptibility.

Chapter 6 Concluding Remarks

ILBC is a breast cancer subtype with distinct clinical and biological features. Although ILBC tumours commonly have features associated with good prognosis, clinicians still face many challenges in the long-term management of women with ILBC. Some of the major challenges relate to the difficulties in detecting ILBC at an early stage, their highly invasive nature and their tendency for distant metastasis. Recent research efforts have demonstrated that ILBC tumours have distinctive genomic (Ciriello *et al.*, 2015), transcriptomic (Zhao *et al.*, 2004; Bertucci *et al.*, 2008) and proteomic (Oliveira *et al.*, 2016) features that has further increased our understanding of this subtype. One of the aims of this thesis was to examine the genome-wide DNA methylation profile of this subtype. The analysis identified 53,898 DMPs between ILBC and non-ILBC tumours that overlapped 13,763 genes and 8,456 intergenic regions. A similar differential methylation profile was observed when the analysis was limited to a subset of luminal A ILBC and luminal A non-ILBC tumours, which suggested that the observed methylation differences were specific to ILBC and non-ILBC tumours and are not influenced by the hormone receptor expression status of the tumours. Many of the differentially methylated genes were found to be involved in biological pathways related to *metabolism of RNA* (R-HSA-8953854), *mRNA processing* (GO:0006397), *RNA splicing* (GO:0008380), *cell cycle* (R-HSA-1640170) and *DNA repair* (GO:0006281). Although this work identified some differences between ILBC and non-ILBC tumours at genome-wide DNA methylation level, further studies are needed to fully understand the impact of such widespread genome-wide DNA methylation changes on gene function.

Investigating the tumour DNA methylation has been shown to be an efficient approach to study tumour heterogeneity (Holm *et al.*, 2010; Fleischer *et al.*, 2017; S Zhang *et al.*, 2018). ILBC is a heterogeneous group of diseases with varying clinical outcomes however, they are often referred to as a single breast cancer subtype, which does not factor in the underlying biological and molecular heterogeneity within ILBC. This generalised approach could limit the opportunity for efficient therapeutic options for women with ILBC. This study exemplified the heterogeneity within ILBC tumours by profiling their genome-wide DNA methylation. Scanning of the ILBC methylome revealed 2,771 regions of variable methylation in ILBC tumours. Replication of the variable methylation analysis in TCGA dataset (ILBC, n=168), identified 2,760 VMRs, of which 763 (28%) overlapped with the study set. The ten most significant VMRs identified in the study set ranked highly in the TCGA dataset. A pooled survival analysis of the study set and TCGA data, after adjustment for age at diagnosis and tumour stage showed that methylation was associated with overall survival for four genes: *APC* (HR = 1.18, 95% CI: 1.02-1.36), *TMEM101* (HR=1.23, 95% CI: 1.02-1.48), *HCG4P3* (HR= 1.37, 95% CI: 1.05-1.79) and *CELF2* (HR=1.21, 95% CI: 1.02-1.43), and promoter methylation at these four genes were identified as potential prognostic biomarkers for women with ILBC. Although, 2,771 VMRs were identified across the genome, only the ten most significant VMRs were tested for their association with survival to minimise the multiple testing burden. It is possible that many other VMRs or individual CpG sites may be associated with survival. This finding provides a base for important future works and could lead to the development of refined molecular signatures for enhanced prediction of ILBC survival with clinical utility, which could not be achieved in the current study due to limited power. These VMRs could also be investigated for their diagnostic potential and if validated in larger studies, could serve in early ILBC diagnosis.

The pioneer work for refining breast cancer classification using gene-expression profiling was largely based on IDBC and only included a small number of ILBC samples (n = 2) (Perou *et al.*, 2000). More research followed that was mainly focused on further refining the gene-expression based intrinsic subtypes of breast cancer. This study provided data to support that homogeneous subgroups of ILBC can be identified using the genome-wide tumour DNA measurement. This approach defined three subgroups of ILBC via unsupervised cluster analysis. It was shown that three subgroups of ILBC (as defined by genome-wide DNA methylation) had significantly (*P*-value < 0.01) different methylation levels at the DMPs, some important differences in epidemiological risk factors and provided some evidence for differences in patient prognosis. One of the important findings of the unsupervised cluster analysis was the identification of ILBC Subgroup 1, which was identified as the most distinct ILBC subgroup defined by a predominantly hypomethylated profiles when compared to the other two subgroups. Tumours from this subgroup clustered alongside the triple-negative non-ILBC cases and were found to be more similar to the TNBC cases in terms of their genome-wide DNA methylation profile compared with the tumours in other two methylation-defined subgroups (difference in global methylation pattern, Subgroup 1 *versus* Subgroup 2, ANNOVA, *P*-value = 3.9×10^{-6}, Subgroup 1 *versus* Subgroup 3, *P*-value = 9.4×10^{-10} ; Subgroup 1 *versus* TNBC, *P*-value = 0.37). Survival analysis using the cox proportional hazard model showed that Subgroup 1 had a poorer overall survival compared with Subgroup 2 (HR: 0.59, 95% CI: 0.19-1.79) and Subgroup 3 (HR: 0.16, 95% CI: 0.03-0.88), after adjusting for age and year of diagnosis that further suggested that Subgroup 1 may represent a more aggressive form of ILBC. Furthermore, the regions of differential methylation between Subgroup 1 and Subgroup 2 showed a significant enrichment of genes involved in immune regulation suggesting that Subgroup 1 may represent a subset

of ILBC tumours with enhanced immune activity as described in previous studies (Ciriello *et al.*, 2015; Michaut *et al.*, 2016). Future work investigating Subgroup 1 identified in this research with more aggressive clinical behaviour (worse prognosis), may identify additional important targets for precision medicine.

Another important finding from the ILBC methylation clustering was the enrichment of Subgroup 3 for cases who had a mother with a history of cancer and both Subgroup 2 and Subgroup 3 for women who had a female relative with a history of breast cancer. This result may indicate that cases in Subgroup 2 and Subgroup 3 could be associated with shared germline genetic variations that may be influencing the tumourigenic pathway sufficiently to generate the defining methylation patterns in these subgroups. This is consistent with the literature that support the association of a strong heritable component with ILBC susceptibility (Allen-Brady *et al.*, 2005; Henry & Cannon-Albright, 2019). Further investigation of the samples from Subgroup 2 and Subgroup 3 for possible association with a shared heritable component could provide important information to support research aimed at identifying subgroup-specific predisposition genes associated with ILBC.

Mutational signatures provide an overview of the processes that could have contributed to the tumour aetiology with utilities for understanding the cause and molecular processes for specific tumour samples. Somatic mutation profiling using WES was performed on a subset of samples from the ILBC methylation-defined subgroups to further characterise the subgroups via mutational signature analysis. MMRd associated signature SBS6 was identified as the most frequently observed mutational signature, detected in 12/15 (80%) of the ILBC tumours. Although the three subgroups could not be

clearly differentiated based on their mutational signature profile, it pointed towards a possible role of MMRd in ILBC tumourigenesis and progression. A computational test checking for microsatellite instability indicated 13/15 (87%) of the ILBC tumours as microsatellite unstable. This is consistent with prior literature that reports a high frequency of microsatellite instability in ILBC tumours (Aldaz *et al.*, 1995; Contegiacomo *et al.*, 1995). The status of mismatch repair genes has been studied extensively in colorectal cancers and to a lesser extent in all breast cancers but the involvement of MMRd in ILBC remains poorly understood. There is growing evidence that mismatch repair deficient tumours benefit from therapies involving immune checkpoint blockade (Le *et al.*, 2015; Le *et al.*, 2017). Genetic and immunohistochemical testing of MMRd are widely available. Future studies on ILBC focusing on the status of mismatch repair genes could identify a subset of ILBC that could benefit from these therapies.

Chapter 7 Bibliography

Agathanggelou A, Cooper WN, & Latif F. (2005). Role of the Ras-association domain family 1 tumor suppressor gene in human cancers. *Cancer research, 65*(9), 3497-3508.

AIHW. (2014). Australian Cancer Incidence and Mortality (ACIM) Books. In: Australian Institute of Health and Welfare Canberra.

Akbari M, Hansen MD, Halgunset J, Skorpen F, & Krokan HE. (2005). Low copy number DNA template can render polymerase chain reaction error prone in a sequence-dependent manner. *The Journal of molecular diagnostics, 7*(1), 36-39.

Al Safar HS, Abidi FH, Khazanehdari KA, Dadour IR, & Tay GK. (2011). Evaluation of different sources of DNA for use in genome wide studies and forensic application. *Appl. Microbiol. Biotechnol., 89*(3), 807-815. doi:10.1007/s00253-010-2926-3

Aldaz CM, Chen T, Sahin A, Cunningham J, & Bondy M. (1995). Comparative allelotype of in situ and invasive human breast cancer: high frequency of microsatellite instability in lobular breast carcinomas. *Cancer research, 55*(18), 3976-3981.

Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, . . . Consortium P. (2020). The repertoire of mutational signatures in human cancer. *Nature, 578*(7793), 94-101. doi:10.1038/s41586-020-1943-3

Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, & Stratton MR. (2013). Deciphering signatures of mutational processes operative in human cancer. *Cell reports, 3*(1), 246-259.

Alexandrov LB, & Stratton MR. (2014). Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. *Current opinion in genetics & development, 24*, 52-60.

Allen-Brady K, Camp NJ, Ward JH, & Cannon-Albright LA. (2005). Lobular breast cancer: excess familiality observed in the Utah Population Database. *Int. J. Cancer, 117*(4), 655-661. doi:10.1002/ijc.21236

Anderson WF, Chatterjee N, Ershler WB, & Brawley OW. (2002). Estrogen receptor breast cancer phenotypes in the Surveillance, Epidemiology, and End Results database. *Breast Cancer Research and Treatment, 76*(1), 27-36.

Anderson WF, Chu KC, Chatterjee N, Brawley O, & Brinton LA. (2001). Tumor variants by hormone receptor expression in white patients with node-negative breast

cancer from the surveillance, epidemiology, and end results database. *Journal of clinical oncology, 19*(1), 18-27.

Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Retrieved from http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Andrieu GP, Shafran JS, Deeney JT, Bharadwaj KR, Rangarajan A, & Denis GV. (2018). BET proteins in abnormal metabolism, inflammation, and the breast cancer microenvironment. *Journal of leukocyte biology, 104*(2), 265-274.

Antoniou A, Pharoah PD, Narod S, Risch HA, Eyfjord JE, Hopper JL, . . . Borg Å. (2003). Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case series unselected for family history: a combined analysis of 22 studies. *The American Journal of Human Genetics, 72*(5), 1117-1130.

Antoniou AC, Casadei S, Heikkinen T, Barrowdale D, Pylkäs K, Roberts J, . . . Fostira F. (2014). Breast-cancer risk in families with mutations in PALB2. *New England Journal of Medicine, 371*(6), 497-506.

Antoniou AC, & Easton D. (2006). Models of genetic susceptibility to breast cancer. *Oncogene, 25*(43), 5898-5905.

Arpino G, Bardou VJ, Clark GM, & Elledge RM. (2004). Infiltrating lobular carcinoma of the breast: tumor characteristics and clinical outcome. *Breast Cancer Res., 6*(3), R149-156. doi:10.1186/bcr767

Arps DP, Healy P, Zhao L, Kleer CG, & Pang JC. (2013). Invasive ductal carcinoma with lobular features: a comparison study to invasive ductal and invasive lobular carcinomas of the breast. *Breast Cancer Research and Treatment, 138*(3), 719-726.

Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, & Irizarry RA. (2014). Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics, 30*(10), 1363-1369. doi:10.1093/bioinformatics/btu049

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, . . . Eppig JT. (2000). Gene ontology: tool for the unification of biology. *Nature genetics, 25*(1), 25-29.

Babiera GV, Lowy AM, Davidson BS, & Singletary SE. (1997). The role of contralateral prophylactic mastectomy in invasive lobular carcinoma. *The breast journal, 3*(1), 2-6.

Bader AG, Kang S, Zhao L, & Vogt PK. (2005). Oncogenic PI3K deregulates transcription and translation. *Nat. Rev. Cancer, 5*(12), 921-929. doi:10.1038/nrc1753

Badve S, Dabbs DJ, Schnitt SJ, Baehner FL, Decker T, Eusebi V, . . . Lakhani SR. (2011). Basal-like and triple-negative breast cancers: a critical review with an emphasis on the implications for pathologists and oncologists. *Modern Pathology, 24*(2), 157-167.

Bae YK, Brown A, Garrett E, Bornman D, Fackler MJ, Sukumar S, . . . Gabrielson E. (2004). Hypermethylation in histologically distinct classes of breast cancer. *Clin. Cancer Res., 10*(18 Pt 1), 5998-6005. doi:10.1158/1078-0432.CCR-04-0667

Bailey CL, Kelly P, & Casey PJ. (2009). Activation of Rap1 promotes prostate cancer metastasis. *Cancer research, 69*(12), 4962-4968.

Banerjee S, Reis-Filho JS, Ashley S, Steele D, Ashworth A, Lakhani SR, & Smith IE. (2006). Basal-like breast carcinomas: clinical outcome and response to chemotherapy. *Journal of clinical pathology, 59*(7), 729-735.

Bardou V-J, Arpino G, Elledge RM, Osborne CK, & Clark GM. (2003). Progesterone receptor status significantly improves outcome prediction over estrogen receptor status alone for adjuvant endocrine therapy in two large breast cancer databases. *Journal of clinical oncology, 21*(10), 1973-1979.

Barker SJ, Anderson E, & Mullen R. (2019). Magnetic resonance imaging for invasive lobular carcinoma: is it worth it? *Gland surgery, 8*(3), 237.

Barreau O, Assié G, Wilmot-Roussel H, Ragazzon B, Baudry C, Perlemoine K, . . . Bertherat J. (2013). Identification of a CpG island methylator phenotype in adrenocortical carcinomas. *J. Clin. Endocrinol. Metab., 98*(1), E174-184. doi:10.1210/jc.2012-2993

Bartlett J, Mallon E, & Cooke T. (2003). The clinical evaluation of HER-2 status: which test to use? *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland, 199*(4), 411-417.

Baylin SB. (2005). DNA methylation and gene silencing in cancer. *Nat. Clin. Pract. Oncol., 2 Suppl 1*, S4-11. doi:10.1038/ncponc0354

Baylln SB, Herman JG, Graff JR, Vertino PM, & Issa J-P. (1997). Alterations in DNA Methylation: A Fundamental Aspect of Neoplasia. In G. F. Vande Woude & G. Klein (Eds.), *Advances in Cancer Research* (Vol. 72, pp. 141-196): Academic Press.

Bediaga NG, Acha-Sagredo A, Guerra I, Viguri A, Albaina C, Diaz IR, . . . Montaner D. (2010). DNA methylation epigenotypes in breast cancer molecular subtypes. *Breast Cancer Research, 12*(5), R77.

Beebe-Dimmer JL, Yee C, Cote ML, Petrucelli N, Palmer N, Bock C, . . . Simon MS. (2015). Familial clustering of breast and prostate cancer and risk of

postmenopausal breast cancer in the W omen's H ealth I nitiative S tudy. *Cancer, 121*(8), 1265-1272.

Bell JT, Pai AA, Pickrell JK, Gaffney DJ, Pique-Regi R, Degner JF, . . . Pritchard JK. (2011). DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome biology, 12*(1), R10.

Ben-Baruch NE, Bose R, Kavuri SM, Ma CX, & Ellis MJ. (2015). HER2-mutated breast cancer responds to treatment with single-agent neratinib, a second-generation HER2/EGFR tyrosine kinase inhibitor. *Journal of the National Comprehensive Cancer Network, 13*(9), 1061-1064.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, . . . Bignell HR. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature, 456*(7218), 53-59.

Beral V, Reeves G, Bull D, Green J, & Collaborators MWS. (2011). Breast cancer risk in relation to the interval between menopause and starting hormone therapy. *Journal of the National Cancer Institute, 103*(4), 296-305.

Bertucci F, Orsetti B, Nègre V, Finetti P, Rougé C, Ahomadegbe J-C, . . . Jacquemier J. (2008). Lobular and ductal carcinomas of the breast have distinct genomic and expression profiles. *Oncogene, 27*(40), 5359-5372.

Berx G, Cleton-Jansen A, Nollet F, De Leeuw W, Van de Vijver M, Cornelisse C, & Van Roy F. (1995). E-cadherin is a tumour/invasion suppressor gene mutated in human lobular breast cancers. *The EMBO journal, 14*(24), 6107-6115.

Bevier M, Sundquist K, & Hemminki K. (2012). Risk of breast cancer in families of multiple affected women and men. *Breast Cancer Research and Treatment, 132*(2), 723-728.

Bevilacqua RA, & Simpson AJ. (2000). Methylation of the hMLH1 promoter but no hMLH1 mutations in sporadic gastric carcinomas with high-level microsatellite instability. *International Journal of Cancer, 87*(2), 200-203.

Bidard F-C, Ng C, Cottu P, Piscuoglio S, Escalup L, Sakr R, . . . Wang L. (2015). Response to dual HER2 blockade in a patient with HER3-mutant metastatic breast cancer. *Annals of Oncology, 26*(8), 1704-1709.

Biglia N, Maggiorotto F, Liberale V, Bounous V, Sgro L, Pecchio S, . . . Ponzone R. (2013). Clinical-pathologic features, long term-outcome and surgical treatment in a large series of patients with invasive lobular carcinoma (ILC) and invasive ductal carcinoma (IDC). *European Journal of Surgical Oncology (EJSO), 39*(5), 455-460.

Bird A. (1992). The essentials of DNA methylation. *Cell, 70*(1), 5-8.

Bird A. (2002). DNA methylation patterns and epigenetic memory. *Genes Dev., 16*(1), 6-21. doi:10.1101/gad.947102

Birth Deaths and Marriages Victoria. (2020). Retrieved from https://www.bdm.vic.gov.au/research-and-family-history/search-your-family-history

Blows FM, Driver KE, Schmidt MK, Broeks A, Van Leeuwen FE, Wesseling J, . . . Blomqvist C. (2010). Subtyping of breast cancer by immunohistochemistry to investigate a relationship between subtype and short and long term survival: a collaborative analysis of data for 10,159 cases from 12 studies. *PLoS med, 7*(5), e1000279.

Bonferroni CE. (1935). Il calcolo delle assicurazioni su gruppi di teste. *Studi in onore del professore salvatore ortu carboni*, 13-60.

Boot A, Huang MN, Ng AW, Ho S-C, Lim JQ, Kawakami Y, . . . Rozen SG. (2018). In-depth characterization of the cisplatin mutational signature in human cell lines and in esophageal and liver tumors. *Genome research, 28*(5), 654-665.

Boyd NF, Dite GS, Stone J, Gunasekara A, English DR, McCredie MR, . . . Yaffe MJ. (2002). Heritability of mammographic density, a risk factor for breast cancer. *New England Journal of Medicine, 347*(12), 886-894.

Brabender J, Usadel H, Danenberg KD, Metzger R, Schneider PM, Lord RV, . . . Danenberg PV. (2001). Adenomatous polyposis coli gene promoter hypermethylation in non-small cell lung cancer is associated with survival. *Oncogene, 20*(27), 3528-3532. doi:10.1038/sj.onc.1204455

Braunstein LZ, Brock JE, Chen Y-H, Truong L, Russo AL, Arvold ND, & Harris JR. (2015). Invasive lobular carcinoma of the breast: local recurrence after breast-conserving therapy by subtype approximation and surgical margin. *Breast Cancer Research and Treatment, 149*(2), 555-564.

Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, & Jemal A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin., 68*(6), 394-424. doi:10.3322/caac.21492

Bray F, McCarron P, & Parkin DM. (2004). The changing global patterns of female breast cancer incidence and mortality. *Breast Cancer Research, 6*(6), 229.

Brem RF, Ioffe M, Rapelyea JA, Yost KG, Weigert JM, Bertrand ML, & Stern LH. (2009). Invasive lobular carcinoma: detection with mammography, sonography, MRI, and breast-specific gamma imaging. *AJR Am. J. Roentgenol., 192*(2), 379-383. doi:10.2214/AJR.07.3827

Brenet F, Moh M, Funk P, Feierstein E, Viale AJ, Socci ND, & Scandura JM. (2011). DNA methylation of the first exon is tightly linked to transcriptional silencing. *PLoS One, 6*(1), e14524. doi:10.1371/journal.pone.0014524

Brooks-Wilson A, Kaurah P, Suriano G, Leach S, Senz J, Grehan N, . . . Bacani J. (2004). Germline E-cadherin mutations in hereditary diffuse gastric cancer: assessment of 42 new families and review of genetic screening criteria. *Journal of medical genetics, 41*(7), 508-517.

Buchegger K, Ili C, Riquelme I, Letelier P, Corvalán AH, Brebi P, . . . Roa JC. (2016). Reprimo as a modulator of cell migration and invasion in the MDA-MB-231 breast cancer cell line. *Biological research, 49*(1), 5.

Buffart TE, Overmeer RM, Steenbergen RD, Tijssen M, van Grieken NC, Snijders PJ, . . . Meijer GA. (2008). MAL promoter hypermethylation as a novel prognostic marker in gastric cancer. *British Journal of Cancer, 99*(11), 1802-1807.

Buoso E, Masi M, Long A, Chiappini C, Travelli C, Govoni S, & Racchi M. (2020). Ribosomes as a nexus between translation and cancer progression: Focus on ribosomal Receptor for Activated C Kinase 1 (RACK1) in breast cancer. *British Journal of Pharmacology*.

Butler R, Venta LA, Wiley E, Ellis R, Dempsey P, & Rubin E. (1999). Sonographic evaluation of infiltrating lobular carcinoma. *AJR. American journal of roentgenology, 172*(2), 325-330.

Caldeira JRF, Prando ÉC, Quevedo FC, Neto FAM, Rainho CA, & Rogatto SR. (2006). CDH1promoter hypermethylation and E-cadherin protein expression in infiltrating breast cancer. *BMC Cancer, 6*(1), 48.

Calza S, Hall P, Auer G, Bjöhle J, Klaar S, Kronenwett U, . . . Smeds J. (2006). Intrinsic molecular signature of breast cancer in a population-based cohort of 412 patients. *Breast Cancer Research, 8*(4), R34.

Cancer Council Australia. (2020). Breast cancer in Australia statistics. Retrieved from https://www.canceraustralia.gov.au/affected-cancer/cancer-types/breast-cancer/statistics

Cancer Genome Atlas N. (2012). Comprehensive molecular portraits of human breast tumours. *Nature, 490*(7418), 61-70. doi:10.1038/nature11412

Cao L, Basudan A, Sikora MJ, Bahreini A, Tasdemir N, Levine KM, . . . Atkinson JM. (2019). Frequent amplifications of ESR1, ERBB2 and MDM4 in primary invasive lobular breast carcinoma. *Cancer Lett., 461*, 21-30. doi:10.1016/j.canlet.2019.06.011

Cao Y, Li Y, Zhang N, Hu J, Yin L, Pan Z, . . . Li F. (2015). Quantitative DNA hypomethylation of ligand Jagged1 and receptor Notch1 signifies occurrence and progression of breast carcinoma. *American journal of cancer research, 5*(5), 1621.

Carey LA, Perou CM, Livasy CA, Dressler LG, Cowan D, Conway K, . . . Edmiston S. (2006). Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study. *JAMA, 295*(21), 2492-2502.

Cawson JN, Law EM, & Kavanagh AM. (2001). Invasive lobular carcinoma: sonographic features of cancers detected in a BreastScreen Program. *Australasian radiology, 45*(1), 25-30.

Chan DS, Abar L, Cariolou M, Nanu N, Greenwood DC, Bandera EV, . . . Norat T. (2019). World Cancer Research Fund International: Continuous Update Project—Systematic literature review and meta-analysis of observational cohort studies on physical activity, sedentary behavior, adiposity, and weight change and breast cancer risk. *Cancer Causes & Control*, 1-18.

Chang X, Monitto CL, Demokan S, Kim MS, Chang SS, Zhong X, . . . Sidransky D. (2010). Identification of hypermethylated genes associated with cisplatin resistance in human cancers. *Cancer research, 70*(7), 2870-2879.

Chapellier C, Balu-Maestro C, Bleuse A, Ettore F, & Bruneton J. (2000). Ultrasonography of invasive lobular carcinoma of the breast: sonographic patterns and diagnostic value: report of 102 cases. *Clinical imaging, 24*(6), 333-336.

Chen Z, Yang J, Li S, Lv M, Shen Y, Wang B, . . . Zhang L. (2017). Invasive lobular carcinoma of the breast: A special histological type compared with invasive ductal carcinoma. *PLoS One, 12*(9), e0182397.

Cheng G, Sun X, Wang J, Xiao G, Wang X, Fan X, . . . Mao Y. (2014). HIC1 silencing in triple-negative breast cancer drives progression through misregulation of LCN2. *Cancer research, 74*(3), 862-872.

Chin K, DeVries S, Fridlyand J, Spellman PT, Roydasgupta R, Kuo W-L, . . . Ryder T. (2006). Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell, 10*(6), 529-541.

Ciriello G, Gatza ML, Beck AH, Wilkerson MD, Rhie SK, Pastore A, . . . Perou CM. (2015). Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. *Cell, 163*(2), 506-519. doi:10.1016/j.cell.2015.09.033

Claus EB, Schildkraut JM, Thompson WD, & Risch NJ. (1996). The genetic attributable risk of breast and ovarian cancer. *Cancer: Interdisciplinary International Journal of the American Cancer Society, 77*(11), 2318-2324.

Clavel-Chapelon F, & Gerber M. (2002). Reproductive factors and breast cancer risk. Do they differ according to age at diagnosis? *Breast Cancer Res. Treat., 72*(2), 107-115.

Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, . . . Noushmehr H. (2016). TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res., 44*(8), e71. doi:10.1093/nar/gkv1507

Collaborative Group on Hormonal Factors in Breast C. (2001). Familial breast cancer: collaborative reanalysis of individual data from 52 epidemiological studies including 58,209 women with breast cancer and 101,986 women without the disease. *Lancet, 358*(9291), 1389-1399. doi:10.1016/S0140-6736(01)06524-2

Collaborative Group on Hormonal Factors in Breast C. (2012). Menarche, menopause, and breast cancer risk: individual participant meta-analysis, including 118 964 women with breast cancer from 117 epidemiological studies. *Lancet Oncol., 13*(11), 1141-1151. doi:10.1016/S1470-2045(12)70425-4

Collaborative Group on Hormonal Factors in Breast Cancer. (2002). Breast cancer and breastfeeding: collaborative reanalysis of individual data from 47 epidemiological studies in 30 countries, including 50 302 women with breast cancer and 96 973 women without the disease. *The Lancet, 360*(9328), 187-195.

Contegiacomo A, Palmirotta R, De Marchis L, Pizzi C, Mastranzo P, Delrio P, . . . Frati L. (1995). Microsatellite instability and pathological aspects of breast cancer. *International Journal of Cancer, 64*(4), 264-268.

Conway K, Edmiston SN, May R, Kuan PF, Chu H, Bryant C, . . . Troester MA. (2014). DNA methylation profiling in the Carolina Breast Cancer Study defines cancer subclasses differing in clinicopathologic characteristics and survival. *Breast Cancer Research, 16*(5), 450.

Costello JF, Frühwald MC, Smiraglia DJ, Rush LJ, Robertson GP, Gao X, . . . Plass C. (2000). Aberrant CpG-island methylation has non-random and tumour-type–specific patterns. *Nat. Genet., 24*(2), 132-138. doi:10.1038/72785

Couto E, & Hemminki K. (2007). Estimates of heritable and environmental components of familial breast cancer using family history information. *British Journal of Cancer, 96*(11), 1740-1742.

Cristofanilli M, Gonzalez-Angulo A, Sneige N, Kau S-W, Broglio K, Theriault RL, . . . Hortobagyi GN. (2005). Invasive lobular carcinoma classic type: response to primary chemotherapy and survival outcomes. *J. Clin. Oncol., 23*(1), 41-48. doi:10.1200/JCO.2005.03.111

Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L, . . . Jassal B. (2010). Reactome: a database of reactions, pathways and biological processes. *Nucleic acids research, 39*(suppl_1), D691-D697.

Curtis C, Shah SP, Chin S-F, Turashvili G, Rueda OM, Dunning MJ, . . . Aparicio S. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature, 486*(7403), 346-352. doi:10.1038/nature10983

Cuzick J, Dowsett M, Pineda S, Wale C, Salter J, Quinn E, . . . Ellis IO. (2011). Prognostic value of a combined estrogen receptor, progesterone receptor, Ki-67, and human epidermal growth factor receptor 2 immunohistochemical score and comparison with the Genomic Health recurrence score in early breast cancer. *J Clin Oncol, 29*(32), 4273-4278.

Dandachi N, Dietze O, & Hauser-Kronberger C. (2002). Chromogenic in situ hybridization: a novel approach to a practical and sensitive method for the detection of HER2 oncogene in archival human breast carcinoma. *Laboratory investigation, 82*(8), 1007-1014.

Dawood S, Hu R, Homes MD, Collins LC, Schnitt SJ, Connolly J, . . . Tamimi RM. (2011). Defining breast cancer prognosis based on molecular phenotypes: results from a large cohort study. *Breast Cancer Research and Treatment, 126*(1), 185-192.

de Almeida BP, Apol√≥nio JD, Binnie A, & Castelo-Branco P. (2019). Roadmap of DNA methylation in breast cancer identifies novel prognostic biomarkers. *BMC Cancer, 19*(1), 219. doi:10.1186/s12885-019-5403-0

Debouki-Joudi S, Trifa F, Khabir A, Sellami-Boudawara T, Frikha M, Daoud J, & Mokdad-Gargouri R. (2017). CpG methylation of APC promoter 1A in sporadic and familial breast cancer patients. *Cancer Biomark., 18*(2), 133-141. doi:10.3233/CBM-160005

Delpech Y, Coutant C, Hsu L, Barranger E, Iwamoto T, Barcenas CH, . . . Pusztai L. (2013). Clinical benefit from neoadjuvant chemotherapy in oestrogen receptor-positive invasive ductal and lobular carcinomas. *British Journal of Cancer, 108*(2), 285-291.

Deng G, Chen A, Hong J, Chae HS, & Kim YS. (1999). Methylation of CpG in a small region of the hMLH1 promoter invariably correlates with the absence of gene expression. *Cancer Res., 59*(9), 2029-2033.

Deniziaut G, Tille JC, Bidard F-C, Vacher S, Schnitzler A, Chemlali W, . . . Rouzier R. (2016). ERBB2 mutations associated with solid variant of high-grade invasive lobular breast carcinomas. *Oncotarget, 7*(45), 73337.

Dent R, Trudeau M, Pritchard KI, Hanna WM, Kahn HK, Sawka CA, . . . Narod SA. (2007). Triple-negative breast cancer: clinical features and patterns of recurrence. *Clinical Cancer Research, 13*(15), 4429-4434.

Derksen PWB, Liu X, Saridin F, van der Gulden H, Zevenhoven J, Evers B, . . . Jonkers J. (2006). Somatic inactivation of E-cadherin and p53 in mice leads to metastatic lobular mammary carcinoma through induction of anoikis resistance and angiogenesis. *Cancer Cell, 10*(5), 437-449. doi:10.1016/j.ccr.2006.09.013

Desmedt C, Zoppoli G, Gundem G, Pruneri G, Larsimont D, Fornili M, . . . Vincent D. (2016). Genomic characterization of primary invasive lobular breast cancer. *Journal of clinical oncology, 34*(16), 1872-1881.

Desmedt C, Zoppoli G, Sotiriou C, & Salgado R. (2017). Transcriptomic and genomic features of invasive lobular breast cancer. *Semin. Cancer Biol., 44*, 98-105. doi:10.1016/j.semcancer.2017.03.007

DiNome ML, Orozco JI, Matsuba C, Manughian-Peter AO, Ensenyat-Mendez M, Chang S-C, . . . Marzese DM. (2019). Clinicopathological Features of Triple-Negative Breast Cancer Epigenetic Subtypes. *Annals of surgical oncology, 26*(10), 3344-3353.

Ditchi Y, Broudin C, El Dakdouki Y, Muller M, Lavaud P, Caron O, . . . Benusiglio PR. (2019). Low risk of invasive lobular carcinoma of the breast in carriers of BRCA1 (hereditary breast and ovarian cancer) and TP53 (Li-Fraumeni syndrome) germline mutations. *Breast J., 25*(1), 16-19. doi:10.1111/tbj.13154

Dixon JM, Anderson TJ, Page DL, Lee D, & Duffy SW. (1982). Infiltrating lobular carcinoma of the breast. *Histopathology, 6*(2), 149-161. doi:10.1111/j.1365-2559.1982.tb02712.x

Dixon JM, Renshaw L, Dixon J, & Thomas J. (2011). Invasive lobular carcinoma: response to neoadjuvant letrozole therapy. *Breast Cancer Research and Treatment, 130*(3), 871-877.

Do H, & Dobrovic A. (2015). Sequence artifacts in DNA from formalin-fixed tissues: causes and strategies for minimization. *Clinical chemistry, 61*(1), 64-71.

Dowsett M, Houghton J, Iden C, Salter J, Farndon J, A'hern R, . . . Baum M. (2006). Benefit from adjuvant tamoxifen therapy in primary breast cancer patients according oestrogen receptor, progesterone receptor, EGF receptor and HER2 status. *Annals of Oncology, 17*(5), 818-826.

Drong AW, Nicholson G, Hedman ÅK, Meduri E, Grundberg E, Small KS, . . . Soranzo N. (2013). The presence of methylation quantitative trait loci indicates a direct genetic influence on the level of DNA methylation in adipose tissue. *PLoS One, 8*(2), e55923.

Droufakou S, Deshmane V, Roylance R, Hanby A, Tomlinson I, & Hart IR. (2001). Multiple ways of silencing E-cadherin gene expression in lobular carcinoma of the breast. *International Journal of Cancer, 92*(3), 404-408. doi:10.1002/ijc.1208

Du M, Su XM, Zhang T, & Xing YJ. (2014). Aberrant promoter DNA methylation inhibits bone morphogenetic protein 2 expression and contributes to drug resistance in breast cancer. *Molecular medicine reports, 10*(2), 1051-1055.

Du P, Zhang X, Huang C-C, Jafari N, Kibbe WA, Hou L, & Lin SM. (2010). Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics, 11*, 587. doi:10.1186/1471-2105-11-587

Dunnwald LK, Rossing MA, & Li CI. (2007). Hormone receptor status, tumor characteristics, and prognosis: a prospective cohort of breast cancer patients. *Breast Cancer Research, 9*(1), R6.

Easton DF, Deffenbaugh AM, Pruss D, Frye C, Wenstrup RJ, Allen-Brady K, . . . Couch FJ. (2007). A systematic genetic assessment of 1,433 sequence variants of unknown clinical significance in the BRCA1 and BRCA2 breast cancer–predisposition genes. *The American Journal of Human Genetics, 81*(5), 873-883.

Easton DF, Pooley KA, Dunning AM, Pharoah PD, Thompson D, Ballinger DG, . . . Luben R. (2007). Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature, 447*(7148), 1087-1093.

Edge SB, Byrd DR, Carducci MA, Compton CC, Fritz A, & Greene F. (2010). *AJCC cancer staging manual* (Vol. 7): Springer New York.

Ehrlich M. (2000). *DNA Alterations in Cancer: Genetic and Epigenetic Changes*: Eaton Publishing Company.

Ehrlich M, Norris KF, Wang RY, Kuo KC, & Gehrke CW. (1986). DNA cytosine methylation and heat-induced deamination. *Bioscience reports, 6*(4), 387-393.

Ellinger J, Bastian PJ, Jurgan T, Biermann K, Kahl P, Heukamp LC, . . . von Ruecker A. (2008). CpG island hypermethylation at multiple gene sites in diagnosis and prognosis of prostate cancer. *Urology, 71*(1), 161-167.

Elston CW, & Ellis IO. (1991). Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology, 19*(5), 403-410.

Ercan C, van Diest PJ, van der Ende B, Hinrichs J, Bult P, Buerger H, . . . Derksen PWB. (2012). p53 mutations in classic and pleomorphic invasive lobular carcinoma of the breast. *Cell. Oncol., 35*(2), 111-118. doi:10.1007/s13402-012-0071-y

Esteller M, Corn PG, Baylin SB, & Herman JG. (2001). A gene hypermethylation profile of human cancer. *Cancer Res., 61*(8), 3225-3229.

Eusebi V, Magalhaes F, & Azzopardi JG. (1992). Pleomorphic lobular carcinoma of the breast: an aggressive tumor showing apocrine differentiation. *Hum. Pathol., 23*(6), 655-662.

Evans N, & Lyons K. (2000). The use of ultrasound in the diagnosis of invasive lobular carcinoma of the breast less than 10 mm in size. *Clinical radiology, 55*(4), 261-263.

Evron E, Umbricht CB, Korz D, Raman V, Loeb DM, Niranjan B, . . . Sukumar S. (2001). Loss of cyclin D2 expression in the majority of breast cancers is associated with promoter hypermethylation. *Cancer research, 61*(6), 2782-2787.

Fackler MJ, McVeigh M, Evron E, Garrett E, Mehrotra J, Polyak K, . . . Argani P. (2003). DNA methylation of RASSF1A, HIN-1, RAR-beta, Cyclin D2 and Twist in in situ and invasive lobular breast carcinoma. *Int. J. Cancer, 107*(6), 970-975. doi:10.1002/ijc.11508

Fackler MJ, McVeigh M, Mehrotra J, Blum MA, Lange J, Lapides A, . . . Sukumar S. (2004). Quantitative multiplex methylation-specific PCR assay for the detection of promoter hypermethylation in multiple genes in breast cancer. *Cancer research, 64*(13), 4442-4452.

Fan Y, Si W, Ji W, Wang Z, Gao Z, Tian R, . . . Zhang F. (2019). Rack1 mediates tyrosine phosphorylation of Anxa2 by Src and promotes invasion and metastasis in drug-resistant breast cancer cells. *Breast Cancer Research, 21*(1), 1-16.

Farmer P, Bonnefoi H, Becette V, Tubiana-Hulin M, Fumoleau P, Larsimont D, . . . Goldstein D. (2005). Identification of molecular apocrine breast tumours by microarray analysis. *Breast Cancer Research, 7*(2), 1-1.

Fechner RE. (1975). Histologic variants of infiltrating lobular carcinoma of the breast. *Human pathology, 6*(3), 373-378.

Fisher ER, Gregorio RM, Redmond C, & Fisher B. (1977). Tubulolobular invasive breast cancer: a variant of lobular invasive cancer. *Human pathology, 8*(6), 679-683.

Fleischer T, Klajic J, Aure MR, Louhimo R, Pladsen AV, Ottestad L, . . . Alnæs GIG. (2017). DNA methylation signature (SAM40) identifies subgroups of the Luminal A breast cancer samples with distinct survival. *Oncotarget, 8*(1), 1074.

Fong Y, Evans J, Brook D, Kenkre J, Jarvis P, & Gower-Thomas K. (2015). The Nottingham Prognostic Index: five-and ten-year data for all-cause survival within a screened population. *Annals of the Royal College of Surgeons of England, 97*(2), 137.

Foote FW, & Stewart FW. (1941). Lobular carcinoma in situ: A rare form of mammary cancer. *Am. J. Pathol., 17*(4), 491-496.493.

Ford D, Easton DF, Bishop DT, Narod SA, & Goldgar DE. (1994). Risks of cancer in BRCA1-mutation carriers. *The Lancet, 343*(8899), 692-695.

Fortin J-P, Labbe A, Lemire M, Zanke BW, Hudson TJ, Fertig EJ, . . . Hansen KD. (2014). Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biol., 15*(12), 503. doi:10.1186/s13059-014-0503-2

Fortunato L, Mascaro A, Poccia I, Andrich R, Amini M, Costarelli L, . . . Vitelli C. (2012). Lobular breast cancer: same survival and local control compared with ductal cancer, but should both be treated the same way? Analysis of an institutional database over a 10-year period. *Annals of surgical oncology, 19*(4), 1107-1114.

Fraga MF, Herranz M, Espada J, Ballestar E, Paz MF, Ropero S, . . . Esteller M. (2004). A mouse skin multistage carcinogenesis model reflects the aberrant DNA methylation patterns of human tumors. *Cancer Res., 64*(16), 5527-5534. doi:10.1158/0008-5472.CAN-03-4061

Fulford LG, Reis-Filho JS, Ryder K, Jones C, Gillett CE, Hanby A, . . . Lakhani SR. (2007). Basal-like grade III invasive ductal carcinoma of the breast: patterns of metastasis and long-term survival. *Breast Cancer Research, 9*(1), R4.

Ghantous A, Saffery R, Cros M-P, Ponsonby A-L, Hirschfeld S, Kasten C, . . . Hernandez-Vargas H. (2014). Optimized DNA extraction from neonatal dried blood spots: application in methylome profiling. *BMC Biotechnol., 14*, 60. doi:10.1186/1472-6750-14-60

Ghoussaini M, & Pharoah PD. (2009). Polygenic susceptibility to breast cancer: current state-of-the-art. *Future oncology, 5*(5), 689-701.

Gibbs JR, van der Brug MP, Hernandez DG, Traynor BJ, Nalls MA, Lai S-L, . . . Troncoso J. (2010). Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genet, 6*(5), e1000952.

Goldhirsch A, Wood WC, Coates AS, Gelber RD, Th√ºrlimann B, J. Senn H, & Panel m. (2011). Strategies for subtypes‚Äîdealing with the diversity of breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2011. *Annals of Oncology, 22*(8), 1736-1747. doi:10.1093/annonc/mdr304

Gozuacik D, Bialik S, Raveh T, Mitou G, Shohat G, Sabanay H, . . . Kimchi A. (2008). DAP-kinase is a mediator of endoplasmic reticulum stress-induced caspase activation and autophagic cell death. *Cell Death & Differentiation, 15*(12), 1875-1886.

Green A, Powe D, Rakha E, Soria D, Lemetre C, Nolan C, . . . Ball G. (2013). Identification of key clinical phenotypes of breast cancer using a reduced panel of protein biomarkers. *British Journal of Cancer, 109*(7), 1886-1894.

Guedj M, Marisa L, De Reynies A, Orsetti B, Schiappa R, Bibeau F, . . . Longy M. (2012). A refined molecular taxonomy of breast cancer. *Oncogene, 31*(9), 1196-1206.

Guilford P, Hopkins J, Harraway J, McLeod M, McLeod N, Harawira P, . . . Reeve AE. (1998). E-cadherin germline mutations in familial gastric cancer. *Nature, 392*(6674), 402-405.

Guo W, Zhu L, Yu M, Zhu R, Chen Q, & Wang Q. (2018). A five-DNA methylation signature act as a novel prognostic biomarker in patients with ovarian serous cystadenocarcinoma. *Clinical epigenetics, 10*(1), 142.

Hall CA, Wang R, Miao J, Oliva E, Shen X, Wheeler T, . . . Goode S. (2010). Hippo pathway effector Yap is an ovarian cancer oncogene. *Cancer research, 70*(21), 8517-8525.

Hao X, Luo H, Krawczyk M, Wei W, Wang W, Wang J, . . . Yi S. (2017). DNA methylation markers for diagnosis and prognosis of common cancers. *Proceedings of the National Academy of Sciences, 114*(28), 7414-7419.

Haque W, Arms A, Verma V, Hatch S, Butler EB, & Teh BS. (2019). Outcomes of pleomorphic lobular carcinoma versus invasive lobular carcinoma. *The Breast, 43*, 67-73.

Harbeck N, Nimmrich I, Hartmann A, Ross JS, Cufer T, Grützmann R, . . . Margossian A. (2008). Multicenter study using paraffin-embedded tumor tissue testing PITX2 DNA methylation as a marker for outcome prediction in tamoxifen-treated, node-negative breast cancer patients. *Journal of clinical oncology, 26*(31), 5036-5042.

Harris JR, Lippman ME, Veronesi U, & Willett W. (1992). Breast cancer. *New England Journal of Medicine, 327*(5), 319-328.

Hassiotou F, & Geddes D. (2013). Anatomy of the human mammary gland: Current status of knowledge. *Clinical anatomy, 26*(1), 29-48.

He H, Argiro L, Dessein H, & Chevillard C. (2007). Improved technique that allows the performance of large-scale SNP genotyping on DNA immobilized by FTA® technology. *Infect. Genet. Evol., 7*(1), 128-132. doi:10.1016/j.meegid.2006.06.001

He K, Zhang L, & Long X. (2016). Quantitative assessment of the association between APC promoter methylation and breast cancer. *Oncotarget, 7*(25), 37920.

Head SR, Komori HK, LaMere SA, Whisenant T, Van Nieuwerburgh F, Salomon DR, & Ordoukhanian P. (2014). Library construction for next-generation sequencing: overviews and challenges. *Biotechniques, 56*(2), 61-77.

Heilat GB, Brennan ME, & French J. (2019). Update on the management of early-stage breast cancer. *Australian journal of general practice, 48*(9), 604.

Hemminki A, Markie D, Tomlinson I, Avizienyte E, Roth S, Loukola A, . . . Höglund P. (1998). A serine/threonine kinase gene defective in Peutz–Jeghers syndrome. *Nature, 391*(6663), 184-187.

Hemminki E, Kennedy DL, Baum C, & Mckinlay SM. (1988). Prescribing of noncontraceptive estrogens and progestins in the United States, 1974-86. *American journal of public health, 78*(11), 1479-1481.

Henrique R, Ribeiro FR, Fonseca D, Hoque MO, Carvalho AL, Costa VL, . . . Jerónimo C. (2007). High promoter methylation levels of APC predict poor prognosis in sextant biopsies from prostate cancer patients. *Clin. Cancer Res., 13*(20), 6122-6129. doi:10.1158/1078-0432.CCR-07-1042

Henry NL, & Cannon-Albright LA. (2019). Breast cancer histologic subtypes show excess familial clustering. *Cancer, 125*(18), 3131-3138.

Herman JG, & Baylin SB. (2003). Gene silencing in cancer in association with promoter hypermethylation. *N. Engl. J. Med., 349*(21), 2042-2054. doi:10.1056/NEJMra023075

Herman JG, Umar A, Polyak K, Graff JR, Ahuja N, Issa J-PJ, . . . Kinzler KW. (1998). Incidence and functional consequences of hMLH1 promoter hypermethylation in colorectal carcinoma. *Proceedings of the National Academy of Sciences, 95*(12), 6870-6875.

Hesson LB, Cooper WN, & Latif F. (2007). The role of RASSF1A methylation in cancer. *Disease markers, 23*(1, 2), 73-87.

Hilleren DJ, Andersson IT, Lindholm K, & Linnell FS. (1991). Invasive lobular carcinoma: mammographic findings in a 10-year experience. *Radiology, 178*(1), 149-154. doi:10.1148/radiology.178.1.1984294

Hoff ER, Tubbs RR, Myles JL, & Procop GW. (2002). HER2/neu amplification in breast cancer: stratification by tumor type and grade. *American journal of clinical pathology, 117*(6), 916-921.

Holm K, Hegardt C, Staaf J, Vallon-Christersson J, Jönsson G, Olsson H, . . . Ringnér M. (2010). Molecular subtypes of breast cancer are associated with characteristic DNA methylation patterns. *Breast Cancer Res., 12*(3), R36. doi:10.1186/bcr2590

Hopper J, Giles G, McCredie M, & Boyle P. (1994). Background, rationale and protocol for a case-control-family study of breast cancer. *The Breast, 3*(2), 79-86.

Horii A, Nakatsuru S, Ichii S, Nagase H, & Nakamura Y. (1993). Multiple forms of the APC gene transcripts and their tissue-specific expression. *Human molecular genetics, 2*(3), 283-287.

Hortobagyi G, Connolly J, D'Orsi C, Edge S, Mittendorf E, Rugo H, . . . Giuliano A. (2017). AJCC cancer staging manual. *Eight Edition*.

Hosford SR, & Miller TW. (2014). Clinical potential of novel therapeutic targets in breast cancer: CDK4/6, Src, JAK/STAT, PARP, HDAC, and PI3K/AKT/mTOR pathways. *Pharmacogenomics and personalized medicine, 7*, 203.

Hsieh P, & Yamane K. (2008). DNA mismatch repair: molecular mechanism, cancer, and ageing. *Mechanisms of ageing and development, 129*(7-8), 391-407.

Hu CY, Mohtat D, Yu Y, Ko Y-A, Shenoy N, Bhattacharya S, . . . Vallumsetla N. (2014). Kidney cancer is characterized by aberrant methylation of tissue-specific enhancers that are prognostic for overall survival. *Clinical Cancer Research, 20*(16), 4349-4360.

Hu X, Wang J, He W, Zhao P, & Ye C. (2018). MicroRNA-433 targets AKT3 and inhibits cell proliferation and viability in breast cancer. *Oncology Letters, 15*(3), 3998-4004.

Hu Z, Fan C, Oh DS, Marron J, He X, Qaqish BF, . . . Dressler L. (2006). The molecular portraits of breast tumors are conserved across microarray platforms. *BMC Genomics, 7*(1), 1-12.

Hua H, Zhang H, Kong Q, & Jiang Y. (2018). Mechanisms for estrogen receptor expression in human cancer. *Experimental hematology & oncology, 7*(1), 1-11.

Hung C-S, Wang S-C, Yen Y-T, Lee T-H, Wen W-C, & Lin R-K. (2018). Hypermethylation of CCND2 in lung and breast cancer is a potential biomarker and drug target. *International journal of molecular sciences, 19*(10), 3096.

Hurtado A, Holmes KA, Ross-Innes CS, Schmidt D, & Carroll JS. (2011). FOXA1 is a key determinant of estrogen receptor function and endocrine response. *Nature genetics, 43*(1), 27-33.

Iorfida M, Maiorano E, Orvieto E, Maisonneuve P, Bottiglieri L, Rotmensz N, . . . Viale G. (2012). Invasive lobular breast cancer: subtypes and outcome. *Breast Cancer Res. Treat., 133*(2), 713-723. doi:10.1007/s10549-012-2002-z

Jia W, Yu T, Cao X, An Q, & Yang H. (2016). Clinical effect of DAPK promoter methylation in gastric cancer. *Medicine, 95*(43), e5040. doi:10.1097/md.0000000000005040

Jiang Y, Cui L, Shen S-h, & Ding L-d. (2012). The prognostic role of RASSF1A promoter methylation in breast cancer: a meta-analysis of published data. *PLoS One, 7*(5), e36780.

Joh JE, Esposito NN, Kiluk JV, Laronga C, Khakpour N, Soliman H, & Catherine Lee M. (2012). Pathologic Tumor Response of Invasive Lobular Carcinoma to Neo-adjuvant Chemotherapy. *The breast journal, 18*(6), 569-574.

John EM, Hopper JL, Beck JC, Knight JA, Neuhausen SL, Senie RT, . . . Breast Cancer Family R. (2004). The Breast Cancer Family Registry: an infrastructure for cooperative multinational, interdisciplinary and translational studies of the genetic epidemiology of breast cancer. *Breast Cancer Res., 6*(4), R375-389. doi:10.1186/bcr801

Kamalakaran S, Varadan V, Russnes HEG, Levy D, Kendall J, Janevski A, . . . Schlichting E. (2011). DNA methylation patterns in luminal breast cancers differ from non-luminal subtypes and can identify relapse risk independent of other clinical variables. *Molecular oncology, 5*(1), 77-92.

Kanehisa M, Furumichi M, Tanabe M, Sato Y, & Morishima K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic acids research, 45*(D1), D353-D361.

Kappil M, Terry MB, Delgado-Cruzata L, Liao Y, & Santella RM. (2016). Mismatch repair polymorphisms as markers of breast cancer prevalence in the breast cancer family registry. *Anticancer research, 36*(9), 4437-4441.

Kaufmann M, Hortobagyi GN, Goldhirsch A, Scholl S, Makris A, Valagussa P, . . . Jonat W. (2006). Recommendations from an international expert panel on the use of neoadjuvant (primary) systemic treatment of operable breast cancer: an update. *Journal of clinical oncology, 24*(12), 1940-1949.

kConFab biospecimen protocol. Protocol for biospecimen collection and processing. Retrieved from https://www.kconfab.org/epidemiology/biospecimen_protocol.html

Kee G-J, Tan RY-C, Rehena S, Lee JJ-X, Zaw MW-W, Lian W-X, . . . Tan BK-T. (2020). Human epidermal growth factor receptor 2 positive rates in invasive lobular breast carcinoma: The Singapore experience. *World Journal of Clinical Oncology, 11*(5), 283-293.

Keinan O, Kedan A, Gavert N, Selitrennik M, Kim S, Karn T, . . . Lev S. (2014). The lipid-transfer protein Nir2 enhances epithelial-mesenchymal transition and facilitates breast cancer metastasis. *Journal of cell science, 127*(21), 4740-4749.

Keller G, Vogelsang H, Becker I, Hutter J, Ott K, Candidus S, . . . Siewert JR. (1999). Diffuse type gastric and lobular breast carcinoma in a familial gastric cancer patient with an E-cadherin germline mutation. *The American Journal of Pathology, 155*(2), 337-342.

Kelsey JL, Gammon MD, & John EM. (1993). Reproductive factors and breast cancer. *Epidemiologic reviews, 15*(1), 36.

Kerkel K, Spadola A, Yuan E, Kosek J, Jiang L, Hod E, . . . Vilain E. (2008). Genomic surveys by methylation-sensitive SNP analysis identify sequence-dependent allele-specific DNA methylation. *Nature genetics, 40*(7), 904-908.

Kharazmi E, Chen T, Narod S, Sundquist K, & Hemminki K. (2014). Effect of multiplicity, laterality, and age at onset of breast cancer on familial risk of breast cancer: a nationwide prospective cohort study. *Breast Cancer Research and Treatment, 144*(1), 185-192.

Kim K, Son M-Y, Jung C-R, Kim D-S, & Cho H-S. (2018). EHMT2 is a metastasis regulator in breast cancer. *Biochemical and biophysical research communications, 496*(2), 758-762.

Kim SJ, Kang H-S, Chang HL, Jung YC, Sim H-B, Lee KS, . . . Lee ES. (2008). Promoter hypomethylation of the N-acetyltransferase 1 gene in breast cancer. *Oncology reports, 19*(3), 663-668.

Kinne DW, Butler JA, Kimmel M, Flehinger BJ, Menendez-Botet C, & Schwartz M. (1987). Estrogen Receptor Protein of Breast Cancer in Patients With Positive Nodes: High Recurrence Rates in the Postmenopausal Estrogen Receptor—Negative Group. *Archives of surgery, 122*(11), 1303-1306.

Kit AH, Nielsen HM, & Tost J. (2012). DNA methylation based biomarkers: practical considerations and applications. *Biochimie, 94*(11), 2314-2337.

Klarenbeek S, Doornebal CW, Kas SM, Bonzanni N, Bhin J, Braumuller TM, . . . Kersten K. (2020). Response of metastatic mouse invasive lobular carcinoma to mTOR inhibition is partly mediated by the adaptive immune system. *Oncoimmunology, 9*(1), 1724049.

Kotsopoulos J, Chen WY, Gates MA, Tworoger SS, Hankinson SE, & Rosner BA. (2010). Risk factors for ductal and lobular breast cancer: results from the nurses' health study. *Breast Cancer Res., 12*(6), R106. doi:10.1186/bcr2790

Krecke KN, & Gisvold JJ. (1993). Invasive lobular carcinoma of the breast: mammographic findings and extent of disease at diagnosis in 184 patients. *AJR Am. J. Roentgenol., 161*(5), 957-960. doi:10.2214/ajr.161.5.8273634

Kryh C, Pietersen C, Rahr H, Christensen R, Wamberg P, & Lautrup M. (2014). Re-resection rates and risk characteristics following breast conserving surgery for breast cancer and carcinoma in situ: a single-centre study of 1575 consecutive cases. *The Breast, 23*(6), 784-789.

Kuchenbaecker KB, Hopper JL, Barnes DR, Phillips K-A, Mooij TM, Roos-Blom M-J, . . . Andrieu N. (2017). Risks of breast, ovarian, and contralateral breast cancer for BRCA1 and BRCA2 mutation carriers. *JAMA, 317*(23), 2402-2416.

Kuismanen SA, Holmberg MT, Salovaara R, de la Chapelle A, & Peltomäki P. (2000). Genetic and epigenetic modification of MLH1 accounts for a major share of microsatellite-unstable colorectal cancers. *The American Journal of Pathology, 156*(5), 1773-1779.

Kulis M, & Esteller M. (2010). DNA methylation and cancer. In *Advances in genetics* (Vol. 70, pp. 27-56): Elsevier.

Kurian AW, Hughes E, Handorf EA, Gutin A, Allen B, Hartman A-R, & Hall MJ. (2017). Breast and ovarian cancer penetrance estimates derived from germline multiple-gene sequencing results in women. *JCO Precision Oncology, 1*, 1-12.

Lai Z, Markovets A, Ahdesmaki M, Chapman B, Hofmann O, McEwen R, . . . Dry JR. (2016). VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res., 44*(11), e108. doi:10.1093/nar/gkw227

Lakhani SR. (2012). *WHO Classification of Tumours of the Breast*: International Agency for Research on Cancer.

Lakhani SR, Ellis IO, Schnitt S, Tan PH, & van de Vijver M. (2012). WHO Classification of Tumours of the Breast.

Lal P, Tan LK, & Chen B. (2005). Correlation of HER-2 status with estrogen and progesterone receptors and histologic features in 3,655 invasive breast carcinomas. *American journal of clinical pathology, 123*(4), 541-546.

Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, . . . Jang W. (2018). ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic acids research, 46*(D1), D1062-D1067.

Lapidus RG, Ferguson AT, Ottaviano YL, Parl FF, Smith HS, Weitzman SA, . . . Davidson NE. (1996). Methylation of estrogen and progesterone receptor gene

5'CpG islands correlates with lack of estrogen and progesterone receptor gene expression in breast tumors. *Clinical Cancer Research, 2*(5), 805-810.

Le DT, Durham JN, Smith KN, Wang H, Bartlett BR, Aulakh LK, . . . Luber BS. (2017). Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. *Science, 357*(6349), 409-413.

Le DT, Uram JN, Wang H, Bartlett BR, Kemberling H, Eyring AD, . . . Laheru D. (2015). PD-1 blockade in tumors with mismatch-repair deficiency. *New England Journal of Medicine, 372*(26), 2509-2520.

Le Gal M, Ollivier L, Asselain B, Meunier M, Laurent M, Vielh P, & Neuenschwander S. (1992). Mammographic features of 455 invasive lobular carcinomas. *Radiology, 185*(3), 705-708. doi:10.1148/radiology.185.3.1438749

Lee JS, Fackler MJ, Lee JH, Choi C, Park MH, Yoon JH, . . . Sukumar S. (2010). Basal-like breast cancer displays distinct patterns of promoter methylation. *Cancer biology & therapy, 9*(12), 1017-1024.

Lee SB, Sohn G, Kim J, Chung IY, Lee JW, Kim HJ, . . . Ahn S-H. (2018). A retrospective prognostic evaluation analysis using the 8th edition of the American Joint Committee on Cancer staging system for breast cancer. *Breast Cancer Research and Treatment, 169*(2), 257-266.

Lehmann U, Celikkaya G, Hasemeier B, Länger F, & Kreipe H. (2002). Promoter hypermethylation of the death-associated protein kinase gene in breast cancer is associated with the invasive lobular subtype. *Cancer Res., 62*(22), 6634-6638.

Lewis MJ, Wiebe JP, & Heathcote JG. (2004). Expression of progesterone metabolizing enzyme genes (AKR1C1, AKR1C2, AKR1C3, SRD5A1, SRD5A2) is altered in human breast carcinoma. *BMC Cancer, 4*(1), 27.

Leygo C, Williams M, Jin HC, Chan MW, Chu WK, Grusch M, & Cheng YY. (2017). DNA methylation as a noninvasive epigenetic biomarker for the detection of cancer. *Disease markers, 2017*.

Li CI, Anderson BO, Daling JR, & Moe RE. (2003). Trends in incidence rates of invasive lobular and ductal breast carcinoma. *JAMA, 289*(11), 1421-1424. doi:10.1001/jama.289.11.1421

Li CI, & Daling JR. (2007). Changes in breast cancer incidence rates in the United States by histologic subtype and race/ethnicity, 1995 to 2004. *Cancer Epidemiol. Biomarkers Prev., 16*(12), 2773-2780. doi:10.1158/1055-9965.EPI-07-0546

Li CI, Uribe DJ, & Daling JR. (2005). Clinical characteristics of different histologic types of breast cancer. *Br. J. Cancer, 93*(9), 1046-1052. doi:10.1038/sj.bjc.6602787

Li H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, . . . Durbin R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics, 25*(16), 2078-2079.

Li J, Yen C, Liaw D, Podsypanina K, Bose S, Wang SI, . . . McCombie R. (1997). PTEN, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer. *Science, 275*(5308), 1943-1947.

Li N, Xie C, & Lu N. (2017). Crosstalk between Hippo signalling and miRNAs in tumour progression. *The FEBS journal, 284*(7), 1045-1055.

Lien H-C, Chen Y-L, Juang Y-L, & Jeng Y-M. (2015). Frequent alterations of HER2 through mutation, amplification, or overexpression in pleomorphic lobular carcinoma of the breast. *Breast Cancer Research and Treatment, 150*(2), 447-455.

Lin Z, Reierstad S, Huang C-C, & Bulun SE. (2007). Novel estrogen receptor-α binding sites and estradiol target genes identified by chromatin immunoprecipitation cloning in breast cancer. *Cancer research, 67*(10), 5017-5024.

Liu Z, Yang L, Cui DX, Liu BL, Zhang XB, Ma WF, & Zhang Q. (2007). Methylation status and protein expression of adenomatous polyposis coli (APC) gene in breast cancer. *Ai Zheng= Aizheng= Chinese Journal of Cancer, 26*(6), 586-590.

Lo P-K, Mehrotra J, D'Costa A, Fackler MJ, Garrett-Mayer E, Argani P, & Sukumar S. (2006). Epigenetic suppression of secreted frizzled related protein 1 (SFRP1) expression in human breast cancer. *Cancer Biol. Ther., 5*(3), 281-286. doi:10.4161/cbt.5.3.2384

Locker A, Ellis I, Elston C, Nicholson R, Robertson J, & Blamey R. (1991). An evaluation of differences in prognosis, recurrence patterns and receptor status between invasive lobular and other invasive carcinomas of the breast. *European journal of surgical oncology: the journal of the European Society of Surgical Oncology and the British Association of Surgical Oncology, 17*(3), 251-257.

Loginov VI, Pronina IV, Burdennyi AM, Pereyaslova EA, Braga EA, Kazubskaya TP, & Kushlinskii NE. (2017). Role of Methylation in the Regulation of Apoptosis Genes APAF1, DAPK1, and BCL2 in Breast Cancer. *Bull. Exp. Biol. Med., 162*(6), 797-800. doi:10.1007/s10517-017-3716-z

Loibl S, Volz C, Mau C, Blohmer J-U, Costa SD, Eidtmann H, . . . Jackisch C. (2014). Response and prognosis after neoadjuvant chemotherapy in 1,051 patients with infiltrating lobular breast carcinoma. *Breast Cancer Research and Treatment, 144*(1), 153-162.

Lopez JK, & Bassett LW. (2009). Invasive lobular carcinoma of the breast: spectrum of mammographic, US, and MR imaging findings. *Radiographics, 29*(1), 165-176. doi:10.1148/rg.291085100

Ma P, Yang X, Kong Q, Li C, Yang S, Li Y, & Mao B. (2014). The ubiquitin ligase RNF220 enhances canonical Wnt signaling through USP7-mediated deubiquitination of β-catenin. *Molecular and cellular biology, 34*(23), 4355-4366.

Ma X-L, Shen M-N, Hu B, Wang B-L, Yang W-J, Lv L-H, . . . Sun Y-F. (2019). CD73 promotes hepatocellular carcinoma progression and metastasis via activating PI3K/AKT signaling by inducing Rap1-mediated membrane localization of P110β and predicts poor prognosis. *Journal of hematology & oncology, 12*(1), 37.

Malik SS, Masood N, Asif M, Ahmed P, Shah ZU, & Khan JS. (2019). Expressional analysis of MLH1 and MSH2 in breast cancer. *Current problems in cancer, 43*(2), 97-105.

Malkin D, Li FP, Strong LC, Fraumeni JF, Nelson CE, Kim DH, . . . Tainsky MA. (1990). Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. *Science, 250*(4985), 1233-1238.

Mann GJ, Thorne H, Balleine RL, Butow PN, Clarke CL, Edkins E, . . . Gattas M. (2006). Analysis of cancer risk and BRCA1 and BRCA2mutation prevalence in the kConFab familial breast cancer resource. *Breast Cancer Research, 8*(1), R12.

Martens JW, Nimmrich I, Koenig T, Look MP, Harbeck N, Model F, . . . Portengen H. (2005). Association of DNA methylation of phosphoserine aminotransferase with response to endocrine therapy in patients with recurrent breast cancer. *Cancer research, 65*(10), 4101-4117.

Martin RM, Middleton N, Gunnell D, Owen CG, & Smith GD. (2005). Breast-feeding and cancer: the Boyd Orr cohort and a systematic review with meta-analysis. *Journal of the National Cancer Institute, 97*(19), 1446-1457.

Martín-Sánchez E, Mendaza S, Ulazia-Garmendia A, Monreal-Santesteban I, Córdoba A, Vicente-García F, . . . Guerrero-Setas D. (2017). CDH22 hypermethylation is an independent prognostic biomarker in breast cancer. *Clinical epigenetics, 9*(1), 7.

Martinez V, & Azzopardi JG. (1979). Invasive lobular carcinoma of the breast: incidence and variants. *Histopathology, 3*(6), 467-488. doi:10.1111/j.1365-2559.1979.tb03029.x

Mas S, Crescenti A, Gassó P, Vidal-Taboada JM, & Lafuente A. (2007). DNA cards: determinants of DNA yield and quality in collecting genetic samples for pharmacogenetic studies. *Basic Clin. Pharmacol. Toxicol., 101*(2), 132-137. doi:10.1111/j.1742-7843.2007.00089.x

Masciari S, Larsson N, Senz J, Boyd N, Kaurah P, Kandel MJ, . . . Miron P. (2007). Germline E-cadherin mutations in familial lobular breast cancer. *Journal of medical genetics, 44*(11), 726-731.

Mathew A, Rajagopal PS, Villgran V, Sandhu GS, Jankowitz RC, Jacob M, . . . Brufsky A. (2017). Distinct pattern of metastases in patients with invasive lobular carcinoma of the breast. *Geburtshilfe und Frauenheilkunde, 77*(6), 660.

Mavaddat N, Antoniou AC, Easton DF, & Garcia-Closas M. (2010). Genetic susceptibility to breast cancer. *Molecular oncology, 4*(3), 174-191.

Mavaddat N, Barrowdale D, Andrulis IL, Domchek SM, Eccles D, Nevanlinna H, . . . Consortium of Investigators of Modifiers of B. (2012). Pathology of breast and ovarian cancers among BRCA1 and BRCA2 mutation carriers: results from the Consortium of Investigators of Modifiers of BRCA1/2 (CIMBA). *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology, 21*(1), 134-147. doi:10.1158/1055-9965.EPI-11-0775

McCredie MR, Dite GS, Giles GG, & Hopper JL. (1998). Breast cancer in Australian women under the age of 40. *Cancer Causes & Control, 9*(2), 189-198.

McRae AF, Powell JE, Henders AK, Bowdler L, Hemani G, Shah S, . . . Montgomery GW. (2014). Contribution of genetic variation to transgenerational inheritance of DNA methylation. *Genome biology, 15*(5), R73.

McSherry EA, Brennan K, Hudson L, Hill AD, & Hopkins AM. (2011). Breast cancer cell migration is regulated through junctional adhesion molecule-A-mediated activation of Rap1 GTPase. *Breast Cancer Research, 13*(2), R31.

Medina-Jaime AD, Reyes-Vargas F, Martinez-Gaytan V, Zambrano-Galvan G, Portillo-DelCampo E, Burciaga-Nava JA, . . . Sifuentes-Alvarez A. (2014). ESR1 and PGR gene promoter methylation and correlations with estrogen and progesterone receptors in ductal and lobular breast cancer. *Asian Pac J Cancer Prev, 15*(7), 3041-3044.

Mei JV, Alexander JR, Adam BW, & Hannon WH. (2001). Use of filter paper for the collection and analysis of human whole blood specimens. *J. Nutr., 131*(5), 1631S-1636S. doi:10.1093/jn/131.5.1631S

Meijers-Heijboer H, van den Ouweland A, Klijn JW, M., de Snoo AO, R., Hollestelle A, Houben M, . . . Seal S. (2002). Low-penetrance susceptibility to breast cancer due to CHEK2* 1100delC in noncarriers of BRCA1 or BRCA2 mutations. *Nature genetics, 31*(1), 55.

Mesman RL, Calléja FM, Hendriks G, Morolli B, Misovic B, Devilee P, . . . Vreeswijk MP. (2019). The functional impact of variants of uncertain significance in BRCA2. *Genetics in Medicine, 21*(2), 293-302.

Mhuircheartaigh JN, Ní Mhuircheartaigh J, Curran C, Hennessy E, & Kerin MJ. (2008). Prospective matched-pair comparison of outcome after treatment for lobular and ductal breast carcinoma. *British Journal of Surgery, 95*(7), 827-833. doi:10.1002/bjs.6042

Michailidou K, Beesley J, Lindstrom S, Canisius S, Dennis J, Lush MJ, . . . Shah M. (2015). Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. *Nature genetics, 47*(4), 373.

Michailidou K, Lindström S, Dennis J, Beesley J, Hui S, Kar S, . . . Rostamianfar A. (2017). Association analysis identifies 65 new breast cancer risk loci. *Nature, 551*(7678), 92.

Michaut M, Chin S-F, Majewski I, Severson TM, Bismeijer T, de Koning L, . . . Bernards R. (2016). Integration of genomic, transcriptomic and proteomic data identifies two biologically distinct subtypes of invasive lobular breast cancer. *Sci. Rep., 6*, 18517. doi:10.1038/srep18517

Middleton LP, Palacios DM, Bryant BR, Krebs P, Otis CN, & Merino MJ. (2000). Pleomorphic lobular carcinoma: morphology, immunohistochemistry, and molecular analysis. *Am. J. Surg. Pathol., 24*(12), 1650-1656.

Mikeska T, Bock C, Do H, & Dobrovic A. (2012). DNA methylation biomarkers in cancer: progress towards clinical implementation. *Expert review of molecular diagnostics, 12*(5), 473-487.

Miki Y, Swensen J, Shattuck-Eidens D, Futreal PA, Harshman K, Tavtigian S, . . . Ding W. (1994). A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science, 266*(5182), 66-71.

Milne R, Fletcher A, MacInnis R, Hodge A, Hopkins A, Bassett J, . . . Jayasekara H. (2017). Cohort profile: the Melbourne collaborative cohort study (health 2020). *International journal of epidemiology, 46*(6), 1757-1757i.

Milne RL, Burwinkel B, Michailidou K, Arias-Perez J-I, Zamora MP, Menéndez-Rodríguez P, . . . Pita G. (2014). Common non-synonymous SNPs associated with breast cancer susceptibility: findings from the Breast Cancer Association Consortium. *Human molecular genetics, 23*(22), 6096-6111.

Mittag F, Kuester D, Vieth M, Peters B, Stolte B, Roessner A, & Schneider-Stock R. (2006). DAPK promotor methylation is an early event in colorectal carcinogenesis. *Cancer Lett., 240*(1), 69-75. doi:10.1016/j.canlet.2005.08.034

Moelans CB, Vlug EJ, Ercan C, Bult P, Buerger H, Cserni G, . . . Derksen PW. (2015). Methylation biomarkers for pleomorphic lobular breast cancer-a short report. *Cellular Oncology, 38*(5), 397-405.

Molland JG, Donnellan M, Janu NC, Carmalt HL, Kennedy CW, & Gillett DJ. (2004). Infiltrating lobular carcinoma--a comparison of diagnosis, management and outcome with infiltrating duct carcinoma. *Breast, 13*(5), 389-396. doi:10.1016/j.breast.2004.03.004

Munsell MF, Sprague BL, Berry DA, Chisholm G, & Trentham-Dietz A. (2014). Body mass index and breast cancer risk according to postmenopausal estrogen-progestin use and hormone receptor status. *Epidemiologic reviews, 36*(1), 114-136.

Murata H, Khattar NH, Gu L, & Li G-M. (2005). Roles of mismatch repair proteins hMSH2 and hMLH1 in the development of sporadic breast cancer. *Cancer letters, 223*(1), 143-150.

Murata H, Khattar NH, Kang Y, Gu L, & Li G-M. (2002). Genetic and epigenetic modification of mismatch repair genes hMSH2 and hMLH1 in sporadic breast cancer with microsatellite instability. *Oncogene, 21*(37), 5696-5703.

Murtagh F, & Legendre P. (2014). Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? *J. Classification, 31*(3), 274-295.

National Breast Cancer Foundation. (2020). Breast cancer anatomy and how cancer starts. Retrieved from https://nbcf.org.au/about-breast-cancer/diagnosis/breast-cancer-anatomy/

Nelson HD, Zakher B, Cantor A, Fu R, Griffin J, O'Meara ES, . . . Trentham-Dietz A. (2012). Risk factors for breast cancer for women aged 40 to 49 years: a systematic review and meta-analysis. *Annals of Internal Medicine, 156*(9), 635-648.

Newcomb PA, Trentham-Dietz A, Hampton JM, Egan KM, Titus-Ernstoff L, Warren Andersen S, . . . Willett WC. (2011). Late age at first full term birth is strongly associated with lobular breast cancer. *Cancer, 117*(9), 1946-1956. doi:10.1002/cncr.25728

Nguyen C, Liang G, Nguyen TT, Tsao-Wei D, Groshen S, Lübbert M, . . . Jones PA. (2001). Susceptibility of nonpromoter CpG islands to de novo methylation in normal and neoplastic cells. *J. Natl. Cancer Inst., 93*(19), 1465-1472. doi:10.1093/jnci/93.19.1465

Nguyen-Dumont T, Mahmoodi M, Hammet F, Tran T, Tsimiklis H, Giles G, . . . Park D. (2015). Hi-Plex targeted sequencing is effective using DNA derived from archival dried blood spots. *Analytical biochemistry, 470*, 48-51.

Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, . . . Stebbings LA. (2012). Mutational processes molding the genomes of 21 breast cancers. *Cell, 149*(5), 979-993.

Nik-Zainal S, Davies H, Staaf J, Ramakrishna M, Glodzik D, Zou X, . . . Wedge DC. (2016). Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature, 534*(7605), 47-54.

Nimmrich I, Sieuwerts AM, Meijer-van Gelder ME, Schwope I, Bolt-de Vries J, Harbeck N, . . . Dietrich D. (2008). DNA hypermethylation of PITX2 is a marker of poor prognosis in untreated lymph node-negative hormone receptor-positive breast cancer patients. *Breast Cancer Research and Treatment, 111*(3), 429-437.

Niu B, Ye K, Zhang Q, Lu C, Xie M, McLellan MD, . . . Ding L. (2014). MSIsensor: microsatellite instability detection using paired tumor-normal sequence data. *Bioinformatics, 30*(7), 1015-1016.

Noushmehr H, Weisenberger DJ, Diefes K, Phillips HS, Pujara K, Berman BP, . . . Cancer Genome Atlas Research N. (2010). Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. *Cancer Cell, 17*(5), 510-522. doi:10.1016/j.ccr.2010.03.017

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, . . . Ako-Adjei D. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic acids research, 44*(D1), D733-D745.

Oliveira NCS, Gomig THB, Milioli HH, Cordeiro F, Costa GG, Urban CA, . . . Ribeiro EMSF. (2016). Comparative proteomic analysis of ductal and lobular invasive breast carcinoma. *Genet. Mol. Res., 15*(2). doi:10.4238/gmr.15027701

Olivotto IA, Bajdik CD, Ravdin PM, Speers CH, Coldman AJ, Norris BD, . . . Gelmon KA. (2005). Population-based validation of the prognostic model ADJUVANT! for early breast cancer. *Journal of clinical oncology, 23*(12), 2716-2725.

Orvieto E, Maiorano E, Bottiglieri L, Maisonneuve P, Rotmensz N, Galimberti V, . . . Viale G. (2008). Clinicopathologic characteristics of invasive lobular carcinoma of the breast: results of an analysis of 530 cases from a single institution. *Cancer, 113*(7), 1511-1520. doi:10.1002/cncr.23811

Paredes J, Albergaria A, Oliveira JT, Jerónimo C, Milanezi F, & Schmitt FC. (2005). P-cadherin overexpression is an indicator of clinical outcome in invasive breast carcinomas and is associated with CDH3 promoter hypomethylation. *Clinical Cancer Research, 11*(16), 5869-5877.

Pedersen MH, Hood BL, Beck HC, Conrads TP, Ditzel HJ, & Leth-Larsen R. (2017). Downregulation of antigen presentation-associated pathway proteins is linked to

poor outcome in triple-negative breast cancer patient tumors. *Oncoimmunology, 6*(5), e1305531.

Pereira B, Chin S-F, Rueda OM, Vollan H-KM, Provenzano E, Bardwell HA, . . . Sammut S-J. (2016). The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. *Nature communications, 7*(1), 1-16.

Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, . . . Botstein D. (2000). Molecular portraits of human breast tumours. *Nature, 406*(6797), 747-752. doi:10.1038/35021093

Pestalozzi BC, Zahrieh D, Mallon E, Gusterson BA, Price KN, Gelber RD, . . . International Breast Cancer Study G. (2008). Distinct clinical and prognostic features of infiltrating lobular carcinoma of the breast: combined results of 15 International Breast Cancer Study Group clinical trials. *J. Clin. Oncol., 26*(18), 3006-3014. doi:10.1200/JCO.2007.14.9336

Peters TJ, Buckley MJ, Statham AL, Pidsley R, Samaras K, V Lord R, . . . Molloy PL. (2015). De novo identification of differentially methylated regions in the human genome. *Epigenetics Chromatin, 8*(1), 6. doi:10.1186/1756-8935-8-6

Pettersson A, Graff RE, Ursin G, dos Santos Silva I, McCormack V, Baglietto L, . . . Chia KS. (2014). Mammographic density phenotypes and risk of breast cancer: a meta-analysis. *Journal of the National Cancer Institute, 106*(5), dju078.

Pfeifer GP, You Y-H, & Besaratinia A. (2005). Mutations induced by ultraviolet light. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis, 571*(1-2), 19-31.

Pharoah PD, Antoniou A, Bobrow M, Zimmern RL, Easton DF, & Ponder BA. (2002). Polygenic susceptibility to breast cancer and implications for prevention. *Nature genetics, 31*(1), 33-36.

Pharoah PD, Day NE, Duffy S, Easton DF, & Ponder BA. (1997). Family history and the risk of breast cancer: a systematic review and meta-analysis. *International Journal of Cancer, 71*(5), 800-809.

Piccart-Gebhart MJ, Procter M, Leyland-Jones B, Goldhirsch A, Untch M, Smith I, . . . Jackisch C. (2005). Trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer. *New England Journal of Medicine, 353*(16), 1659-1672.

Ping Z, Siegal GP, Harada S, Eltoum I-E, Youssef M, Shen T, . . . Li Y. (2016). ERBB2 mutation is associated with a worse prognosis in patients with CDH1 altered invasive lobular cancer of the breast. *Oncotarget, 7*(49), 80655.

Piqué L, de Paz AM, Piñeyro D, Martínez-Cardús A, de Moura MC, Llinàs-Arias P, . . . Sigurdsson S. (2019). Epigenetic inactivation of the splicing RNA-binding protein CELF2 in human breast cancer. *Oncogene, 38*(45), 7106-7112.

Poon SL, Pang S-T, McPherson JR, Yu W, Huang KK, Guan P, . . . Heng HL. (2013). Genome-wide mutational signatures of aristolochic acid and its application as a screening tool. *Science translational medicine, 5*(197), 197ra101-197ra101.

Porter AJ, Evans EB, Foxcroft LM, Simpson PT, & Lakhani SR. (2014). Mammographic and ultrasound features of invasive lobular carcinoma of the breast. *Journal of Medical Imaging and Radiation Oncology, 58*(1), 1-10. doi:10.1111/1754-9485.12080

Prat A, Parker JS, Karginova O, Fan C, Livasy C, Herschkowitz JI, . . . Perou CM. (2010). Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Res., 12*(5), R68. doi:10.1186/bcr2635

Primakoff P, & Myles DG. (2000). The ADAM gene family: surface proteins with adhesion and protease activity. *Trends Genet., 16*(2), 83-87. doi:10.1016/s0168-9525(99)01926-5

Qian Q, Lv Y, & Li P. (2018). SOCS1 is associated with clinical progression and acts as an oncogenic role in triple-negative breast cancer. *IUBMB Life, 70*(4), 320-327.

Qin Y, Feng H, Chen M, Wu H, & Zheng X. (2018). InfiniumPurify: an R package for estimating and accounting for tumor purity in cancer methylation research. *Genes & diseases, 5*(1), 43-45.

Quach N, Goodman MF, & Shibata D. (2004). In vitro mutation artifacts after formalin fixation and error prone translesion synthesis during PCR. *BMC clinical pathology, 4*(1), 1.

Radpour R, Kohler C, Haghighi M, Fan A, Holzgreve W, & Zhong X. (2009). Methylation profiles of 22 candidate genes in breast cancer using high-throughput MALDI-TOF mass array. *Oncogene, 28*(33), 2969-2978.

Rahman N, Seal S, Thompson D, Kelly P, Renwick A, Elliott A, . . . Chagtai T. (2007). PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. *Nature genetics, 39*(2), 165-167.

Rakha EA, El-Sayed ME, Green AR, Paish EC, Powe DG, Gee J, . . . Ellis IO. (2007). Biologic and clinical characteristics of breast cancer with single hormone receptor–positive phenotype. *Journal of clinical oncology, 25*(30), 4772-4778.

Rakha EA, El-Sayed ME, Menon S, Green AR, Lee AHS, & Ellis IO. (2008). Histologic grading is an independent prognostic factor in invasive lobular carcinoma of the

breast. *Breast Cancer Res. Treat., 111*(1), 121-127. doi:10.1007/s10549-007-9768-4

Rakha EA, & Ellis IO. (2010). *Lobular breast carcinoma and its variants.* Paper presented at the Seminars in diagnostic pathology.

Rakha EA, Gill MS, El-Sayed ME, Khan MM, Hodi Z, Blamey RW, . . . Ellis IO. (2009). The biological and clinical characteristics of breast carcinoma with mixed ductal and lobular morphology. *Breast Cancer Research and Treatment, 114*(2), 243-250.

Rakha EA, & Green AR. (2017). Molecular classification of breast cancer: what the pathologist needs to know. *Pathology, 49*(2), 111-119. doi:10.1016/j.pathol.2016.10.012

Rakha EA, Reis-Filho JS, Baehner F, Dabbs DJ, Decker T, Eusebi V, . . . Lakhani SR. (2010). Breast cancer prognostic classification in the molecular era: the role of histological grade. *Breast Cancer Research, 12*(4), 1-12.

Rakha EA, Reis-Filho JS, & Ellis IO. (2010). Combinatorial biomarker expression in breast cancer. *Breast Cancer Research and Treatment, 120*(2), 293-308.

Reed AEM, Kutasovic JR, Lakhani SR, & Simpson PT. (2015). Invasive lobular carcinoma of the breast: morphology, biomarkers and'omics. *Breast Cancer Res., 17*(1), 12.

Reeves GK, Beral V, Green J, Gathani T, Bull D, & Million Women Study C. (2006). Hormonal therapy for menopause and breast-cancer risk by histological type: a cohort study and meta-analysis. *Lancet Oncol., 7*(11), 910-918. doi:10.1016/S1470-2045(06)70911-1

Renwick A, Thompson D, Seal S, Kelly P, Chagtai T, Ahmed M, . . . Spanova K. (2006). ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. *Nature genetics, 38*(8), 873-875.

Richiardi L, Fiano V, Vizzini L, De Marco L, Delsedime L, Akre O, . . . Merletti F. (2009). Promoter methylation in APC, RUNX3, and GSTP1 and mortality inprostate cancer patients.

Roessler J, Ammerpohl O, Gutwein J, Steinemann D, Schlegelberger B, Weyer V, . . . Schmutzler R. (2015). The CpG island methylator phenotype in breast cancer is associated with the lobular subtype. *Epigenomics, 7*(2), 187-199.

Rosa-Rosa JM, Caniego-Casas T, Leskela S, Cristobal E, González-Martínez S, Moreno-Moreno E, . . . Garrido P. (2019). High frequency of ERBB2 activating mutations in invasive lobular breast carcinoma with pleomorphic features. *Cancers, 11*(1), 74.

Rosenthal R, McGranahan N, Herrero J, Taylor BS, & Swanton C. (2016). DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol., 17*, 31. doi:10.1186/s13059-016-0893-4

Ross JS, Wang K, Sheehan CE, Boguniewicz AB, Otto G, Downing SR, . . . Ali S. (2013). Relapsed classic E-Cadherin (CDH1)–mutated invasive lobular breast cancer shows a high frequency of HER2 (ERBB2) gene mutations. *Clinical Cancer Research, 19*(10), 2668-2676.

Roy R, Chun J, & Powell SN. (2012). BRCA1 and BRCA2: different roles in a common pathway of genome protection. *Nature Reviews Cancer, 12*(1), 68-78.

Sailer V, Gevensleben H, Dietrich J, Goltz D, Kristiansen G, Bootz F, & Dietrich D. (2017). Clinical performance validation of PITX2 DNA methylation as prognostic biomarker in patients with head and neck squamous cell carcinoma. *PLoS One, 12*(6).

Saimura M, Fukutomi T, Tsuda H, Sato H, Miyamoto K, Akashi-Tanaka S, & Nanasawa T. (1999). Prognosis of a series of 763 consecutive node-negative invasive breast cancer patients without adjuvant therapy: Analysis of clinicopathological prognostic factor. *Journal of Surgical Oncology, 71*(2), 101-105.

Salvesen HB, MacDonald N, Ryan A, Iversen OE, Jacobs IJ, Akslen LA, & Das S. (2000). Methylation of hMLH1 in a population-based series of endometrial carcinomas. *Clinical Cancer Research, 6*(9), 3607-3613.

Sarrió D, Moreno-Bueno G, Hardisson D, Sánchez-Estévez C, Guo M, Herman JG, . . . Palacios J. (2003). Epigenetic and genetic alterations of APC and CDH1 genes in lobular breast cancer: relationships with abnormal E-cadherin and catenin expression and microsatellite instability. *International Journal of Cancer, 106*(2), 208-215.

Satterthwaite FE. (1946). An approximate distribution of estimates of variance components. *Biometrics bulletin, 2*(6), 110-114.

Savage P, Yu N, Dumitra S, & Meterissian S. (2019). The effect of the American Joint Committee on Cancer eighth edition on breast cancer staging and prognostication. *European Journal of Surgical Oncology, 45*(10), 1817-1820.

Sawyer E, Roylance R, Petridis C, Brook MN, Nowinski S, Papouli E, . . . Garcia-Closas M. (2014). Genetic predisposition to in situ and invasive lobular carcinoma of the breast. *PLoS Genet., 10*(4), e1004285. doi:10.1371/journal.pgen.1004285

Schelfout K, Van Goethem M, Kersschot E, Verslegers I, Biltjes I, Leyman P, . . . Gillardin J. (2004). Preoperative breast MRI in patients with invasive lobular breast cancer. *European radiology, 14*(7), 1209-1216.

Schrader K, Masciari S, Boyd N, Salamanca C, Senz J, Saunders D, . . . Tung N. (2011). Germline mutations in CDH1 are infrequent in women with early-onset or familial lobular breast cancers. *Journal of medical genetics, 48*(1), 64-68.

Schuster-Böckler B, & Lehner B. (2012). Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature, 488*(7412), 504-507.

Seniski GG, Camargo AA, Ierardi DF, Ramos EAS, Grochoski M, Ribeiro ESF, . . . Klassen G. (2009). ADAM33 gene silencing by promoter hypermethylation as a molecular marker in breast invasive lobular carcinoma. *BMC Cancer, 9*, 80. doi:10.1186/1471-2407-9-80

Sharma G, Mirza S, Parshad R, Srivastava A, Gupta SD, Pandya P, & Ralhan R. (2010). CpG hypomethylation of MDR1 gene in tumor and serum of invasive ductal breast carcinoma patients. *Clinical biochemistry, 43*(4-5), 373-379.

Sharma S, Barry M, O'Reilly E, & Kell M. (2015). Surgical management of lobular carcinoma from a national screening program: a retrospective analysis. *European Journal of Surgical Oncology (EJSO), 41*(1), 79-85.

Shaw J, Walsh T, Chappell S, Carey N, Johnson K, & Walker R. (1996). Microsatellite instability in early sporadic breast cancer. *British Journal of Cancer, 73*(11), 1393-1397.

Shawky SA, El-Borai MH, Khaled HM, Guda I, Mohanad M, Abdellateif MS, . . . Bahanasy AA. (2019). The prognostic impact of hypermethylation for a panel of tumor suppressor genes and cell of origin subtype on diffuse large B-cell lymphoma. *Mol. Biol. Rep., 46*(4), 4063-4076. doi:10.1007/s11033-019-04856-x

Sheng X, Guo Y, & Lu Y. (2017). Prognostic role of methylated GSTP1, p16, ESR1 and PITX2 in patients with breast cancer: A systematic meta-analysis under the guideline of PRISMA. *Medicine, 96*(28).

Shousha S, Backhous CM, Alaghband-Zadeh J, & Burn I. (1986). Alveolar variant of invasive lobular carcinoma of the breast: a tumor rich in estrogen receptors. *American journal of clinical pathology, 85*(1), 1-5.

Silberfein EJ, Hunt KK, Broglio K, Shen J, Sahin A, Le-Petross H, . . . Mittendorf EA. (2010). Clinicopathologic factors associated with involved margins after breast-conserving surgery for invasive lobular carcinoma. *Clinical breast cancer, 10*(1), 52-58.

Singh P, Yang M, Dai H, Yu D, Huang Q, Tan W, . . . Shen B. (2008). Overexpression and hypomethylation of flap endonuclease 1 gene in breast and other cancers. *Molecular Cancer Research, 6*(11), 1710-1717.

Sisti JS, Collins LC, Beck AH, Tamimi RM, Rosner BA, & Eliassen AH. (2016). Reproductive risk factors in relation to molecular subtypes of breast cancer: Results from the nurses' health studies. *International Journal of Cancer, 138*(10), 2346-2356.

Slamon DJ, Clark GM, Wong SG, Levin WJ, Ullrich A, & McGuire WL. (1987). Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science, 235*(4785), 177-182.

Slamon DJ, Leyland-Jones B, Shak S, Fuchs H, Paton V, Bajamonde A, . . . Pegram M. (2001). Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. *New England Journal of Medicine, 344*(11), 783-792.

Smith ZD, Chan MM, Mikkelsen TS, Gu H, Gnirke A, Regev A, & Meissner A. (2012). A unique regulatory phase of DNA methylation in the early mammalian embryo. *Nature, 484*(7394), 339-344. doi:10.1038/nature10960

Smyth GK. (2005). limma: Linear Models for Microarray Data. In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* (pp. 397-420): Springer, New York, NY.

Son KS, Kang H-S, Kim SJ, Jung S-Y, Min SY, Lee SY, . . . Shin KH. (2010). Hypomethylation of the interleukin-10 gene in breast cancer tissues. *The Breast, 19*(6), 484-488.

Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, . . . Jeffrey SS. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences, 98*(19), 10869-10874.

Spurdle AB, Healey S, Devereau A, Hogervorst FB, Monteiro AN, Nathanson KL, . . . Wappenschmidt B. (2012). ENIGMA—evidence-based network for the interpretation of germline mutant alleles: An international initiative to evaluate risk and clinical significance associated with sequence variation in BRCA1 and BRCA2 genes. *Human mutation, 33*(1), 2-7.

Stefansson OA, Moran S, Gomez A, Sayols S, Arribas-Jorba C, Sandoval J, . . . Jonasson JG. (2015). A DNA methylation-based definition of biologically distinct breast cancer subtypes. *Molecular oncology, 9*(3), 555-568.

Stirzaker C, Zotenko E, Song JZ, Qu W, Nair SS, Locke WJ, . . . Dobrovic A. (2015). Methylome sequencing in triple-negative breast cancer reveals distinct methylation clusters with prognostic value. *Nature communications, 6*(1), 1-11.

Stone AL, Kroeger M, & Sang QXA. (1999). Structure,ÄìFunction Analysis of the ADAM Family of Disintegrin-Like and Metalloproteinase-Containing Proteins (Review). *J. Protein Chem., 18*(4), 447-465. doi:10.1023/A:1020692710029

Stouffer SA, Suchman EA, DeVinney LC, Star SA, & Williams Jr RM. (1949). The american soldier: Adjustment during army life.(studies in social psychology in world war ii), vol. 1.

Suriano G, Yew S, Ferreira P, Senz J, Kaurah P, Ford JM, . . . Young S. (2005). Characterization of a recurrent germ line mutation of the E-cadherin gene: implications for genetic testing and clinical management. *Clinical Cancer Research, 11*(15), 5401-5409.

Sutcliffe S, Pharoah PD, Easton DF, Ponder BA, & Group UFOCS. (2000). Ovarian and breast cancer risks to women in families with two or more cases of ovarian cancer. *International Journal of Cancer, 87*(1), 110-117.

Takeichi M. (1991). Cadherin cell adhesion receptors as a morphogenetic regulator. *Science, 251*(5000), 1451-1455.

Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, . . . Dawson E. (2019). COSMIC: the catalogue of somatic mutations in cancer. *Nucleic acids research, 47*(D1), D941-D947.

Taylor DL, Jackson AU, Narisu N, Hemani G, Erdos MR, Chines PS, . . . Welch RP. (2019). Integrative analysis of gene expression, DNA methylation, physiological traits, and genetic variation in human skeletal muscle. *Proceedings of the National Academy of Sciences, 116*(22), 10883-10888.

Thawani JP, Wang AC, Than KD, Lin C-Y, La Marca F, & Park P. (2010). Bone morphogenetic proteins and cancer: review of the literature. *Neurosurgery, 66*(2), 233-246.

Therneau T. (2014). A package for survival analysis in S. R package version 2.37-7.

Thomas G, Jacobs KB, Kraft P, Yeager M, Wacholder S, Cox DG, . . . Yu K. (2009). A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11. 2 and 14q24. 1 (RAD51L1). *Nature genetics, 41*(5), 579-584.

Toyota M, Ahuja N, Ohe-Toyota M, Herman JG, Baylin SB, & Issa J-PJ. (1999). CpG island methylator phenotype in colorectal cancer. *Proceedings of the National Academy of Sciences, 96*(15), 8681-8686.

Toyota M, & Suzuki H. (2010). Epigenetic drivers of genetic alterations. In *Advances in genetics* (Vol. 70, pp. 309-323): Elsevier.

Triche TJ, Jr., Weisenberger DJ, Van Den Berg D, Laird PW, & Siegmund KD. (2013). Low-level processing of Illumina Infinium DNA Methylation BeadArrays. *Nucleic Acids Res., 41*(7), e90. doi:10.1093/nar/gkt090

Tserga A, Michalopoulos NV, Levidou G, Korkolopoulou P, Zografos G, Patsouris E, & Saetta AA. (2012). Association of aberrant DNA methylation with clinicopathological features in breast cancer. *Oncology reports, 27*(5), 1630-1638.

Tubiana-Hulin M, Stevens D, Lasry S, Guinebretiere J, Bouita L, Cohen-Solal C, . . . Rouesse J. (2006). Response to neoadjuvant chemotherapy in lobular and ductal breast carcinomas: a retrospective study on 860 patients from one institution. *Annals of Oncology, 17*(8), 1228-1233.

Turnbull C, Ahmed S, Morrison J, Pernet D, Renwick A, Maranian M, . . . Healey CS. (2010). Genome-wide association study identifies five new breast cancer susceptibility loci. *Nature genetics, 42*(6), 504-507.

Van't Veer LJ, Dai H, Van De Vijver MJ, He YD, Hart AA, Mao M, . . . Witteveen AT. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature, 415*(6871), 530-536.

Victorian Cancer Registry. (2020). Breast cancer case registration. Retrieved from https://registry.cancervic.org.au/registry

Virmani AK, Rathi A, Sathyanarayana UG, Padar A, Huang CX, Cunnigham HT, . . . Gilcrease M. (2001). Aberrant methylation of the adenomatous polyposis coli (APC) gene promoter 1A in breast and lung carcinomas. *Clinical Cancer Research, 7*(7), 1998-2004.

Vos MD, Ellis CA, Bell A, Birrer MJ, & Clark GJ. (2000). Ras uses the novel tumor suppressor RASSF1 as an effector to mediate apoptosis. *Journal of Biological Chemistry, 275*(46), 35669-35672.

Vuong D, Simpson PT, Green B, Cummings MC, & Lakhani SR. (2014). Molecular classification of breast cancer. *Virchows Archiv, 465*(1), 1-14.

Waddington, & H C. (1942). The epigenotype. *Endeavour, 1*, 18-20.

Wang AT, Vachon CM, Brandt KR, & Ghosh K. (2014). *Breast density and breast cancer risk: a practical review.* Paper presented at the Mayo Clinic Proceedings.

Wang S, Li H, Wang J, Wang D, Yao A, & Li Q. (2014). Prognostic and biological significance of microRNA-127 expression in human breast cancer. *Disease markers, 2014*.

Wang W, Dong Y, Li X, Pan Y, Du J, & Liu D. (2020). MicroRNA-431 serves as a tumor inhibitor in breast cancer through targeting FGF9. *Oncology Letters, 19*(1), 1001-1007.

Wang Y, Weng X, Wang L, Hao M, Li Y, Hou L, . . . Lin G. (2018). HIC1 deletion promotes breast cancer progression by activating tumor cell/fibroblast crosstalk. *The Journal of clinical investigation, 128*(12), 5235-5250.

Wasif N, Maggard MA, Ko CY, & Giuliano AE. (2010). Invasive lobular vs. ductal breast cancer: a stage-matched comparison of outcomes. *Annals of surgical oncology, 17*(7), 1862-1869.

Watermann DO, Tempfer C, Hefler LA, Parat C, & Stickeler E. (2005). Ultrasound morphology of invasive lobular breast cancer is different compared with other types of breast cancer. *Ultrasound in medicine & biology, 31*(2), 167-174.

Weidner N, & Semple JP. (1992). Pleomorphic variant of invasive lobular carcinoma of the breast. *Human pathology, 23*(10), 1167-1171.

Weigelt B, Horlings H, Kreike B, Hayes M, Hauptmann M, Wessels L, . . . Peterse J. (2008). Refinement of breast cancer classification by molecular characterization of histological special types. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland, 216*(2), 141-150.

Weiss A, Chavez-MacGregor M, Lichtensztajn DY, Yi M, Tadros A, Hortobagyi GN, . . . Mittendorf EA. (2018). Validation study of the American Joint Committee on Cancer eighth edition prognostic stage compared with the anatomic stage in breast cancer. *JAMA oncology, 4*(2), 203-209.

West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, . . . Nevins JR. (2001). Predicting the clinical status of human breast cancer by using gene expression profiles. *Proceedings of the National Academy of Sciences, 98*(20), 11462-11467.

Widschwendter M, & Jones PA. (2002). DNA methylation and breast carcinogenesis. *Oncogene, 21*(35), 5462-5482. doi:10.1038/sj.onc.1205606

Widschwendter M, Siegmund KD, Müller HM, Fiegl H, Marth C, Müller-Holzner E, . . . Laird PW. (2004). Association of breast cancer DNA methylation profiles with hormone receptor status and response to tamoxifen. *Cancer research, 64*(11), 3807-3813.

Wilson AS, Power BE, & Molloy PL. (2007). DNA hypomethylation and human diseases. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer, 1775*(1), 138-162.

Wirapati P, Sotiriou C, Kunkel S, Farmer P, Pradervand S, Haibe-Kains B, . . . Schütz F. (2008). Meta-analysis of gene expression profiles in breast cancer: toward a

unified understanding of breast cancer subtyping and prognosis signatures. *Breast Cancer Research, 10*(4), R65.

Wise LA, Titus-Ernstoff L, Newcomb PA, Trentham-Dietz A, Trichopoulos D, Hampton JM, & Egan KM. (2009). Exposure to breast milk in infancy and risk of breast cancer. *Cancer Causes & Control, 20*(7), 1083-1090.

Wolff AC, Hammond M, Schwartz JN, Hagerty KL, Allred DC, Cote RJ, . . . Langer A. (2007). College of American Pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer. *J Clin Oncol, 25*(1), 118-145.

Wong EM, Joo JE, McLean CA, Baglietto L, English DR, Severi G, . . . Giles GG. (2015). Tools for translational epigenetic studies involving formalin-fixed paraffin-embedded human tissue: applying the Infinium HumanMethyation450 Beadchip assay to large population-based studies. *BMC research notes, 8*(1), 543.

Wong EM, Joo JE, McLean CA, Baglietto L, English DR, Severi G, . . . Milne RL. (2016). Analysis of the breast cancer methylome using formalin-fixed paraffin-embedded tumour. *Breast Cancer Research and Treatment, 160*(1), 173-180.

Wooster R, Bignell G, Lancaster J, Swift S, Seal S, Mangion J, . . . Micklem G. (1995). Identification of the breast cancer susceptibility gene BRCA2. *Nature, 378*(6559), 789-792.

World Health Organization S, Jack A, World Health O, Percy C, Shanmugarathan S, & Whelan S. (2000). *International Classification of Diseases for Oncology: ICD-O*: World Health Organization.

Wu L, Wang F, Xu R, Zhang S, Peng X, Feng Y, . . . Lu C. (2013). Promoter methylation of BRCA1 in the prognosis of breast cancer: a meta-analysis. *Breast Cancer Research and Treatment, 142*(3), 619-627.

Xia B, Shan M, Wang J, Zhong Z, Geng J, He X, . . . Pang D. (2017). Homeobox A11 hypermethylation indicates unfavorable prognosis in breast cancer. *Oncotarget, 8*(6), 9794.

Xiao B, Chen L, Ke Y, Hang J, Cao L, Zhang R, . . . Chen J. (2018). Identification of methylation sites and signature genes with prognostic value for luminal breast cancer. *BMC Cancer, 18*(1), 405.

Xie ZM, Li LS, Laquet C, Penault-Llorca F, Uhrhammer N, Xie XM, & Bignon YJ. (2011). Germline mutations of the E-cadherin gene in families with inherited invasive lobular breast carcinoma but no diffuse gastric cancer. *Cancer, 117*(14), 3112-3117.

Xing L, Tang X, Wu K, Huang X, Yi Y, & Huan J. (2020). TRIM27 Functions as a Novel Oncogene in Non-Triple-Negative Breast Cancer by Blocking Cellular Senescence through p21 Ubiquitination. *Molecular Therapy-Nucleic Acids, 22*, 910-923.

Xu X, Gammon MD, Zhang Y, Cho YH, Wetmur JG, Bradshaw PT, . . . Neugut AI. (2010). Gene promoter methylation is associated with increased mortality among women with breast cancer. *Breast Cancer Research and Treatment, 121*(3), 685-692.

Xu Y, Diao L, Chen Y, Liu Y, Wang C, Ouyang T, . . . Fan T. (2013). Promoter methylation of BRCA1 in triple-negative breast cancer predicts sensitivity to adjuvant chemotherapy. *Annals of Oncology, 24*(6), 1498-1505.

Yaghjyan L, Mahoney M, Succop P, Wones R, Buckholz J, & Pinney S. (2012). Relationship between breast cancer risk factors and mammographic breast density in the Fernald Community Cohort. *British Journal of Cancer, 106*(5), 996-1003.

Yan M, Li X, Tong D, Han C, Zhao R, He Y, & Jin X. (2016). miR-136 suppresses tumor invasion and metastasis by targeting RASAL2 in triple-negative breast cancer. *Oncology reports, 36*(1), 65-71.

Yan PS, Perry MR, Laux DE, Asare AL, Caldwell CW, & Huang TH-M. (2000). CpG island arrays: an application toward deciphering epigenetic signatures of breast cancer. *Clinical Cancer Research, 6*(4), 1432-1438.

Yang J, Mani SA, Donaher JL, Ramaswamy S, Itzykson RA, Come C, . . . Weinberg RA. (2004). Twist, a master regulator of morphogenesis, plays an essential role in tumor metastasis. *Cell, 117*(7), 927-939.

Yee CJ, Roodi N, Verrier CS, & Parl FF. (1994). Microsatellite instability and loss of heterozygosity in breast cancer. *Cancer research, 54*(7), 1641-1644.

Yin M, & Zhang L. (2011). Hippo signaling: a hub of growth control, tumor suppression and pluripotency maintenance. *Journal of Genetics and Genomics, 38*(10), 471-481.

Yoshinaka T, Nishii K, Yamada K, Sawada H, Nishiwaki E, Smith K, . . . Higashiyama S. (2002). Identification and characterization of novel mouse and human ADAM33s with potential metalloprotease activity. *Gene, 282*(1-2), 227-236. doi:10.1016/s0378-1119(01)00818-6

You JS, & Jones PA. (2012). Cancer genetics and epigenetics: two sides of the same coin? *Cancer Cell, 22*(1), 9-20.

Zhang H, Ahearn TU, Lecarpentier J, Barnes D, Beesley J, Qi G, . . . Bolla MK. (2020). Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. *Nature genetics*, 1-10.

Zhang N, Zuo L, Zheng H, Li G, & Hu X. (2018). Increased expression of CD81 in breast cancer tissue is associated with reduced patient prognosis and increased cell migration and proliferation in MDA-MB-231 and MDA-MB-435S human breast cancer cell lines in vitro. *Medical science monitor: international medical journal of experimental and clinical research, 24*, 5739.

Zhang S, Wang Y, Gu Y, Zhu J, Ci C, Guo Z, . . . Liu H. (2018). Specific breast cancer prognosis-subtype distinctions based on DNA methylation patterns. *Molecular oncology, 12*(7), 1047-1060.

Zhao H, Langer√∏d A, Ji Y, Nowels KW, Nesland JM, Tibshirani R, . . . Jeffrey SS. (2004). Different gene expression patterns in invasive lobular and ductal carcinomas of the breast. *Mol. Biol. Cell, 15*(6), 2523-2536. doi:10.1091/mbc.e03-11-0786

Zheng W, Long J, Gao Y-T, Li C, Zheng Y, Xiang Y-B, . . . Haines JL. (2009). Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25. 1. *Nature genetics, 41*(3), 324-328.

Zheng Y, Shao X, Huang Y, Shi L, Chen B, Wang X, . . . Zhang X. (2016). Role of estrogen receptor in breast cancer cell gene expression. *Molecular medicine reports, 13*(5), 4046-4050.

Zhong Z, Shan M, Wang J, Liu T, Xia B, Niu M, . . . Pang D. (2015). HOXD13 methylation status is a prognostic indicator in breast cancer. *International journal of clinical and experimental pathology, 8*(9), 10716.

Zhou Y, Chen J, Li Q, Huang W, Lan H, & Jiang H. (2015). Association between breastfeeding and breast cancer risk: evidence from a meta-analysis. *Breastfeeding medicine, 10*(3), 175-182.

Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, . . . Chanda SK. (2019). Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat. Commun., 10*(1), 1523. doi:10.1038/s41467-019-09234-6

Zhu S, Ward BM, Yu J, Matthew-Onabanjo AN, Janusis J, Hsieh C-C, . . . Kandil D. (2018). IRS2 mutations linked to invasion in pleomorphic invasive lobular carcinoma. *JCI insight, 3*(8).

Zou D, Yoon H-S, Perez D, Weeks RJ, Guilford P, & Humar B. (2009a). Epigenetic silencing in non-neoplastic epithelia identifies E-cadherin (CDH1) as a target for

chemoprevention of lobular neoplasia. *J. Pathol., 218*(2), 265-272. doi:10.1002/path.2541

Zou D, Yoon HS, Perez D, Weeks RJ, Guilford P, & Humar B. (2009b). Epigenetic silencing in non-neoplastic epithelia identifies E-cadherin (CDH1) as a target for chemoprevention of lobular neoplasia. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland, 218*(2), 265-272.

Chapter 8 Appendices

Additional information 1: R Scripts for data analyses

Differentially methylated positions (DMPs)

```
# load the required R package
library (limma)

# Create a factor object with information about the comparison groups.
group <- factor (pheno$group, levels=c ("group 1", "group 2"))

# Create a design matrix that contains a coefficient for group comparison.
design <- model.matrix (~ group)

# Make pair-wise comparison
fit <- lmFit (M-value, design)
fit <- eBayes (fit)

# Extract a table of the top-ranked genes from a linear model fit.
topTable (fit, coef, number = nrow (M-value), sort.by= "B", adjust.method= "bonferroni", p.value = 0.01)
```

where,

pheno: A data frame with sample information.

M-value: A matrix of M-values, with unique Illumina probe ID as rownames and unique sample IDs as column names.

fit: list containing a linear model fit produced by lmFit.

coef: Column name specifying which coefficient or contrast of the linear model is of interest.

Differentially methylated regions (DMRs)

```
# load the required R package
library (limma)
library (DMRcate)

# Create a factor object with information about the comparison groups.
group <- factor (pheno$group, levels = c (“group 1”, “group 2”))


# Create a design matrix that contains a coefficient for group comparison.
design <- model.matrix (~ 0+ group)

# Construct the contrast matrix corresponding to specified contrasts of a set of parameters.
cont.matrix <- makeContrasts (group 1vsgroup 2 = group 1- group 2, levels = design)

# Annotate a matrix representing 450K data with probe weights
DMR <- cpg.annotate (datatype = "array", M-value, what = "M", arraytype = "450K", analysis.type = "differential", design = design, contrasts = TRUE, cont.matrix = cont.matrix, fdr = 0.01, coef = " group 1vsgroup 2")

# Computes a kernel estimate against a null comparison to identify significantly differentially methylated regions.
dmrcate (DMR, lambda = 1000, C = 2)

# Takes a dmrcate.output object and produces the corresponding GRanges object.
range <- extractRanges (DMR, genome = "hg19")
results <- DMR$results
grange <- data.frame (seqnames = seqnames (range), starts = start (range), ends = end (range), strands = strand (range), no.cpgs = range$no.cpgs, Promoters = range$overlapping.promoter)

# Generate a data frame of differentially methylated regions.
DMR <- cbind (results, grange)
```

where,

pheno: A data frame with sample phenotype information.

M-value: A matrix of M-values, with unique Illumina probe ID as rownames and unique sample IDs as column names.

coef: Column name specifying which coefficient or contrast of the linear model is of interest.

lambda: Gaussian kernel bandwidth for smoothed-function estimation.

C: Scaling factor for bandwidth. Gaussian kernel is calculated where lambda/C = sigma.

Variably methylated regions (VMRs)

```
# load the required R package
library (DMRcate)

# Annotate a matrix representing 450K data with probe weights
VMR <- cpg.annotate (datatype = "array", M-value, what = "M", arraytype = "450K", analysis.type = "variability", contrasts = FALSE, cont.matrix = NULL)

# Computes a kernel estimate against a null comparison to identify significantly variably methylated regions.
VMR <- dmrcate (VMR, lambda = 1000, C = 2)
range <- extractRanges (VMR, genome = "hg19")
results <- VMR$results
grange <- data.frame (seqnames = seqnames (range), starts = start (range), ends = end (range), strands = strand (range), no.cpgs = range$no.cpgs, Promoters = range$overlapping.promoter)

# Generate a data frame of variably methylated regions.
VMR <- cbind (results, grange)
```

where,

M-value: A matrix of M-values, with unique Illumina probe ID as rownames and unique sample IDs as column names.

lambda: Gaussian kernel bandwidth for smoothed-function estimation.

C: Scaling factor for bandwidth. Gaussian kernel is calculated where lambda/C = sigma.

Survival analysis

load the required R package
library (*survival*)
library (*survminer*)

Create a survival object
pheno$SurvObj <- with (pheno, Surv(time, status = = 1))

# Create survival curves from either a formula (e.g. the Kaplan-Meier), a previously fitted Cox model, or a previously fitted accelerated failure time model.
km.by.cluster <- survfit (SurvObj ~ Subgroup, data = pheno, conf.type = "log-log")

where,

pheno: a data frame in which to interpret the variables named in the formula.

SurvObj: A survival object, usually used as a response variable in a model formula.

time: For right censored data, this is the follow up time. For interval data, the first argument is the starting time for the interval.

Subgroup: Comparison groups.

status: The status indicator, 0=alive, 1=dead.

km.by.cluster: An object of class survfit containing one or more survival curves.

Unsupervised cluster analysis

# Compute and return the distance matrix computed by using the specified distance measure to compute the distances between the rows of a data matrix.
dist.matrix <- dist (as.matrix (t (M-value)))

# Hierarchical cluster analysis on a set of dissimilarities and methods for analysing it.
hc <- hclust (dist.matrix, method = "*ward.D2*")

where,

dist.matrix: the distance matrix computed by using Euclidean measure to compute the distances between the rows of a data matrix.

hc: An object of class hclust which describes the tree produced by the clustering process.

M-value: A matrix of M-values, with unique Illumina probe ID as rownames and unique sample IDs as column names.

method: the agglomeration method to be used for clustering.

Mutation signature analysis

```
# Load the required R package
library(deconstructSigs)

# Read the VCF file
vcf.to.sigs.input ("file.vcf")

# Given a mutation list, outputs a data frame with counts of how frequently a mutation is found within each trinucleotide context per sample ID.
sigs.input <- mut.to.sigs.input (mut.ref = , sample.id = "sample", chr = "chr", pos = "pos", ref = "ref", alt = "alt")

# Load a reference signature
reference.signatures <- as.data.frame (t(cancer_signatures_v3))

# Compute mutational signature
whichSignature (tumor.ref= sigs.input, signature.ref = Reference.signatures.cosmic)
```

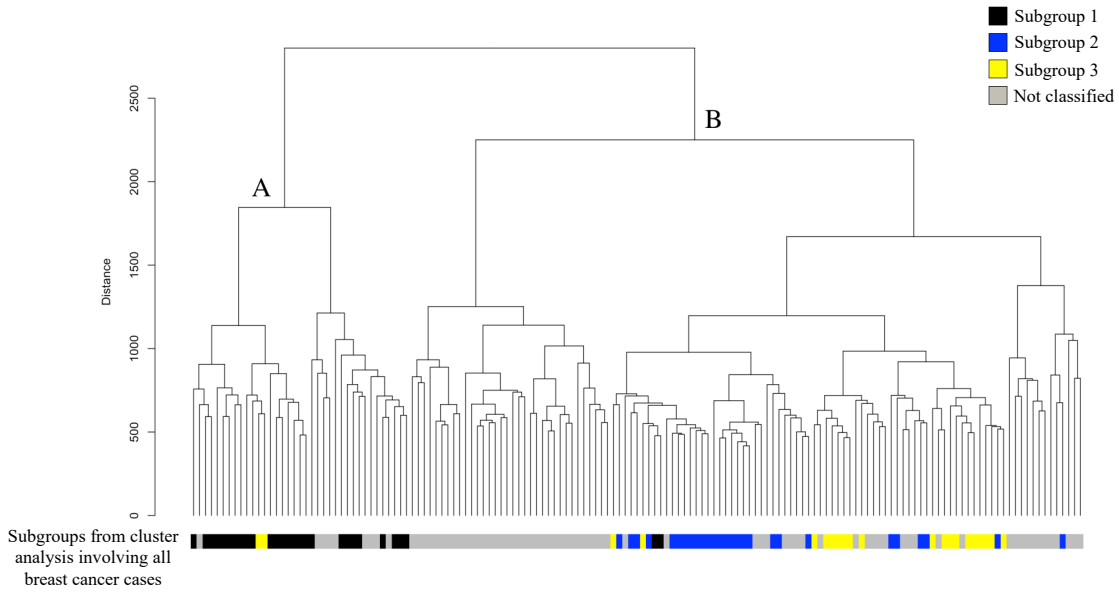
where,

file = location of VCF file that is to be converted.

mut.ref = location of the mutation file that is to be converted.

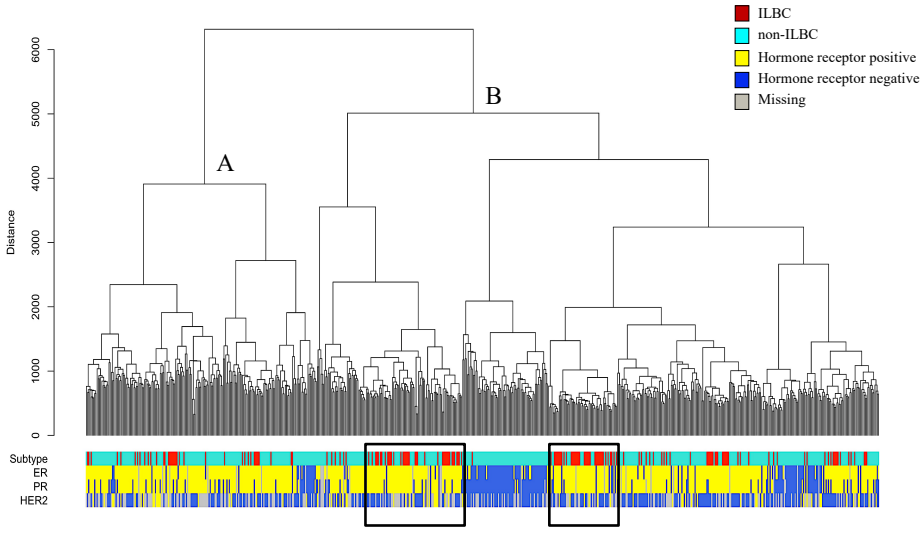
Additional Figure 1: Unsupervised cluster analysis of ILBC samples (n=151) based on their genome-wide DNA methylation profiles.

Dendrogram showing the unsupervised clustering of ILBC cases (n=151) based on their genome-wide DNA methylation levels at 449,005 CpG positions. Each leaf of the dendrogram represent a ILBC sample and the length of the branches show the Euclidean distance between the two clusters. Higher distance on the dendrogram represents more dissimilar clusters and vice-versa. The colour bar "Subgroups from cluster analysis involving all breast cancer cases" indicates the ILCB methylation subgroups identified in the cluster analysis including all breast cancer samples (ILBC, n=151 and non-ILBC, n=341). Subgroup 1, Subgroup 2 and Subgroup 3 are shown in black, violet and yellow colour respectively, on the colour bar. Samples that were not classified into any subgroup in the clustering involving all breast cancer are shown in grey.



Additional Figure 1: Unsupervised cluster analysis of TCGA breast cancer samples (n=666) based on their genome-wide DNA methylation profiles.

Dendrogram showing the unsupervised clustering of TCGA breast cancer samples (n = 666) including ILBC, n = 171 and non-ILBC, n = 495, based on their genome-wide DNA methylation profiles. Each leaf of the dendrogram represents a breast cancer sample and the length of the branches show the Euclidean distance between the two clusters (y-axis). Higher distance on the dendrogram represents more dissimilar clusters and vice-versa. The colour bar "Subtype" indicates the two breast cancer histological subtypes; i.e., ILBC (shown in red) and non-ILBC (shown in turquoise). The colour bars "ER", "PR" and "HER2" indicate the estrogen, progesterone, and human epidermal growth factor receptor 2 expression status, respectively of the breast tumours. Hormone receptor positive tumours are shown in "yellow" and hormone receptor negative tumours are shown in "blue" colour. Black boxes show the samples with similar clustering pattern as methylation-defined Subgroup 1.



Additional Table 1: Whole-exome sequencing quality metrics for the tumour and germline samples included in the pilot study.

Sample	DNA insert size (bp)	Overlapping reads (%)	Mean mapping quality
Tumour			
Sample 1-FFPE	139	44	60
Sample 1-FFPE-rep	176	40	60
Sample 2-FFPE	141	44	60
Sample 2-FFPE-rep	133	45	60
Germline			
Sample 1-WB	202	31	60
Sample 1-GC	173	37	60
Sample 2-WB	188	34	60
Sample 2-GC	179	36	60

Additional File 1: Familial Aspects of Cancer Conference, Kingscliff, New South Wales, Australia 2017- Poster presentation.

PEDIGREE
Pathology, Epidemiology, DNA, Informatics and Genetics: A Research Enabling Enterprise

THE UNIVERSITY OF MELBOURNE · MONASH University · Cancer Council Victoria

http://www.pedigree.org.au/

Invasive Lobular Breast Cancer : Using tumour genome-wide DNA methylation for subtyping and aid in the identification of susceptibility genes

Medha Suman[1], JiHoo Eric Joo[1,2], Ee Ming Wong[1,2], Tu Nguyen-Dumont[1,2], Neil O'Callaghan[1], Melissa Yow[1], ABCFS[3], MCCS[4], kConFab[5], John L. Hopper[3], Graham G. Giles[3, 4, 6], Roger Milne[3,4], Melissa C. Southey[1,2,4] .

[1] Genetic Epidemiology Laboratory, Department of Pathology, Victorian Comprehensive Cancer Centre, The University of Melbourne, [2] Precision Medicine, School of Clinical Sciences at Monash Health, Monash University, [3]Centre for Biostatistics and Epidemiology, School of Population and Global Health, The University of Melbourne, [4]Cancer Epidemiology and Intelligence Division, Cancer Council Victoria, [5]Peter MacCallum Cancer Centre, Victorian Comprehensive Cancer Centre, Melbourne, [6] School of Public Health and Preventive Medicine, Monash University

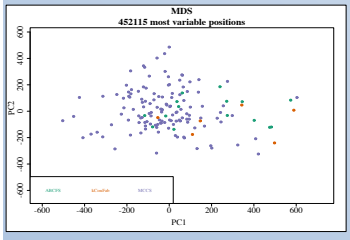
Background

- Invasive Lobular Breast Cancer (ILC) accounts for 10-15% of all breast cancer cases.
- Studies suggest that there is a strong familial risk associated with ILC, which is significantly greater than any other breast cancer subtype [1].
- A whole-genome sequencing project has been conducted in our lab involving 120 early onset and multiple-case ILC families.
- The extent of genetic variation in these genomes made interpretation challenging.
- **The aim of this project is to assess the genome-wide DNA methylation pattern of ILC to enable further subtyping and interpretation of the germline genomic data.**

Samples and Method

- Samples (n=152) in this project were sourced from :-
  - ➢ Melbourne Collaborative Cohort Study (**MCCS**)
  - ➢ Australian Breast Cancer Family Registry (**ABCFR**)
  - ➢ Kathleen Cuningham Foundation Consortium for research into Familial Breast cancer (**kConFab**)
- Formalin-fixed paraffin-embedded (FFPE) tumor-enriched DNA was prepared using macrodissection and run on the Infinium HumanMethylation450K Beadchip array to generate a genome-wide methylation data at ~450,000 methylation sites.
- The raw intensity data was preprocessed and normalized in R programming software using *minfi*, a Bioconductor package for analysis of DNA methylation microarray data [2].

Results



- Multidimensional scaling (MDS) was first performed to investigate overall methylation similarities between individual samples (**Figure 1**).

Figure 1: Overall methylation similarities between samples.

- **E-Cadherin** : Approximately 70% of ILC show a complete loss of E-cadherin protein, average methylation across CDH1 was assessed for all 152 samples. Hypomethylated promoter regions across most ILC tumours (**Figure 2**) was found.

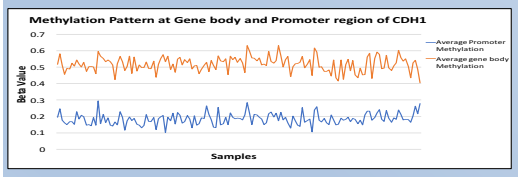
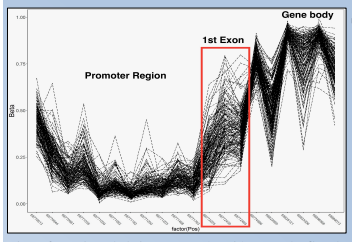


Figure 2: Methylation at CDH1 promoter and gene body.



- A small domain (~3 CpG probes) in CDH1 had highly variable methylation (**Figure3**). ENCODE data suggest regulatory function for this short region

Figure 3: DNA methylation patterns (x-axis) across the CDH1 gene (20 probes)

Results (Cont.)

- We identified variably methylated regions (VMRs) between the lobular breast cancers using DMRcate [3]. **(Table 1)**

Table1: List of top 10 variably methylated regions (VMRs).

Gene Promoters	CHR location	Gene Description	Min FDR
APC	Chr 5	Adenomatous polyposis coli, WNT signaling pathway regulator	6.69E-154
TMEM101	Chr 17	Transmembrane protein 101, NF-kappa-B signalling pathway	3.88E-112
HIST3H2A	Chr 1	Histone Cluster 3 H2A, nucleosome structure maintenance	3.18E-106
ISM1	Chr 20	Isthmin1(novel endogenous angiogenesis inhibitor)	1.87E-97
ASCL2	Chr 11	Achaete-Scute Family BHLH Transcription Factor 2, Human Early Embryo Development	1.33E-91
NKX6-2	Chr 10	NK6 homeobox 2, transcription factor activity	5.75E-83
CCDC108	Chr 2	Coiled-coil domain containing 108,	8.88E-81
DNAJB6	Chr 7	DnaJ heat shock protein family (Hsp40) member B6, protein folding	3.26E-79
SGCE	Chr 7	sarcoglycan epsilon, Paternally expressed imprinted gene	2.35E-77
SPPL2B	Chr 19	Signal peptide peptidase like 2B, immune response	2.45E-77

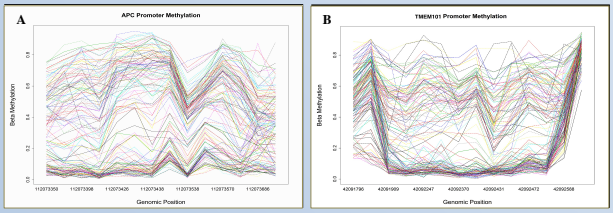


Figure 4: Methylation pattern at the two most variably methylated regions across the APC promoter (A) and TMEM101 (B).

- DNA methylation of two other APC interacting genes (**JUP**, **AXIN1**).

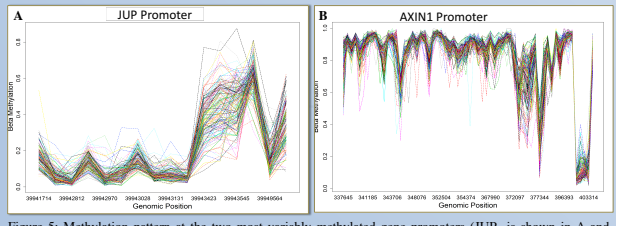


Figure 5: Methylation pattern at the two most variably methylated gene promoters (JUP is shown in A and AXIN1 shown in (B).

Conclusion

- Our genome-wide DNA methylation data suggests that there is large heterogeneity within the invasive lobular breast cancer subtype. This heterogeneity is observed across multiple regions including key tumour suppressor genes such as APC and CDH1 (exon 1).

References:

1. Allen-Brady *et. al* (2005). International Journal of Cancer. 117 (4): 655–61.
2. Aryee MJ *et. al* (2014) Bioinformatics, 30(10), pp. 1363–1369.
3. Peters TJ *et. al* (2015) Epigenetics & Chromatin, 8, pp. 6.

Additional File 2: Familial Aspects of Cancer Conference, Kingscliff, New South Wales, Australia 2018- Poster presentation.

PEDIGREE
Pathology, Epidemiology, DNA, Informatics and Genetics: A Research Enabling Enterprise

VICTORIAN COMPREHENSIVE CANCER CENTRE | THE UNIVERSITY OF MELBOURNE | MONASH University | Cancer Council Victoria

http://www.pedigree.org.au/

Invasive Lobular Breast Cancer : Using tumor genome-wide DNA methylation for subtyping and aid in the identification of susceptibility genes

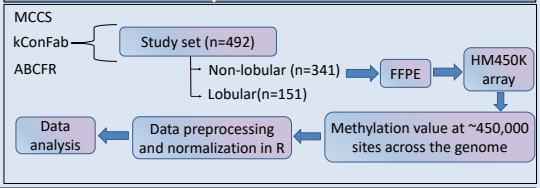
Medha Suman[1], JiHoon Eric Joo[1], Ee Ming Wong[1,2], Tu Nguyen-Dumont[1,2], Neil O'Callaghan[1], Melissa Yow[1], ABCFS[3], MCCS[4], kConFab[5], John L. Hopper[3], Graham G. Giles[3, 4, 6], Roger Milne[2,3,4], Melissa C. Southey[1,2,4] .

[1]Genetic Epidemiology Laboratory, Department of Clinical Pathology, Victorian Comprehensive Cancer Centre, The University of Melbourne, [2] Precision Medicine, School of Clinical Sciences at Monash Health, Monash University, [3]Centre for Epidemiology and Biostatistics, School of Population and Clinical Health, The University of Melbourne, [4]Cancer Epidemiology and Intelligence Division, Cancer Council Victoria, [5]Peter MacCallum Cancer Centre, Victorian Comprehensive Cancer Centre, Melbourne, [6]School of Public Health and Preventive Medicine, Monash University

Background

- Invasive Lobular Breast Cancer (ILC) is a distinct subtype of breast cancer and accounts for 10-15% of all breast cancer cases.
- A whole-genome sequencing project has been conducted in our lab involving 120 early onset and multiple-case ILC families. The extent of genetic variation in these genomes made interpretation challenging [1].
- **The aim of this project is to subtype ILC based on variation in tumour DNA methylation and to aid in the interpretation of the germline genomic data.**

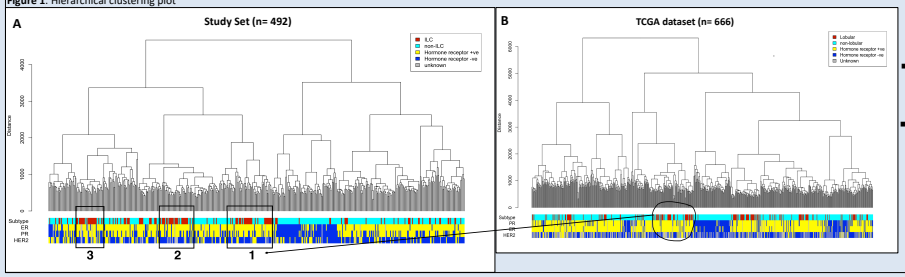
Samples and Method



Results

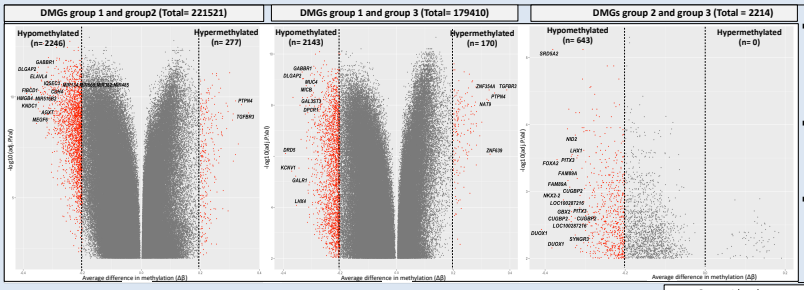
- Unsupervised clustering analysis was performed on all breast cancer samples based on their DNA methylation level at 449,005 CpG sites using "ward.D2" minimum variance method (Fig 1A). Exactly the same analysis was performed on the TCGA breast cancer samples.

Figure 1: Hierarchical clustering plot



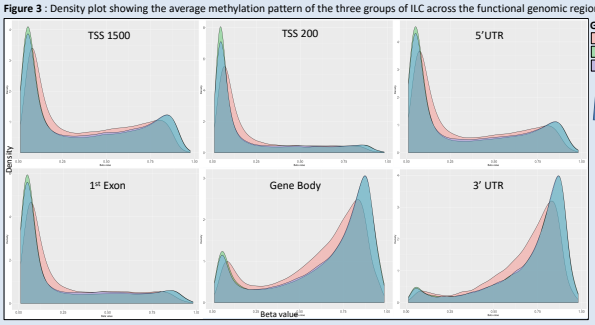
- **A**- Lobular samples were found to cluster into three main groups (group1, group2 and group3).
- **B**- A very similar clustering pattern was observed in TCGA breast cancer samples.
- An overall survival analysis was performed to see if there is any association between the difference in methylation and the overall survival between the groups. However, no significant association was observed.

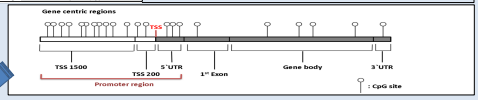
Figure 2 : Volcano plots showing the DMGs between the three groups.



- We performed a pair-wise differential methylation analysis using *limma* [2] linear fit model to identify the differentially methylated genes (DMGs) between the three identified groups of ILC (Fig 2).
- The DMGs with < 20% average methylation difference (Δβ) between the groups were considered insignificant and were removed from further analysis.
- The DMGs with Δβ=positive were defined as hypermethylated genes and those with Δβ=negative were defined as hypomethylated genes.

Figure 3 : Density plot showing the average methylation pattern of the three groups of ILC across the functional genomic regions





Conclusion and future directions

- Our study identified three groups of lobular breast cancer based on their genome-wide DNA methylation pattern.
- DMG analysis identified group 1 as the most hypomethylated group and group 3 as the most hypermethylated group.
- Survival analysis does not indicate any significant association between the difference in methylation and the overall survival. However, small study size might be a reason.
- We propose to do somatic whole exome sequencing on the methylation groups with an aim to identify any potential difference in their mutation profile.
- We also propose do expression profiling on the top DMGs to see if the methylation has any potential effect on the gene expression.

References
1. Cancer Australia – Grant No. APP1047347 (2013-2016), High Risk Genes for Lobular Breast Cancer. Southey MC, Goldgar DE, Park DJ, Lonie A, Winship I, Simpon P, Lakhani S.
2. Peters TJ et. al (2015). "De novo identification of differentially methylated regions in the human genome." Epigenetics & Chromatin, 8, pp. 6.

Additional File 3: Lorne Cancer Conference, Lorne, Australia 2018 -Poster presentation.

Lorne Genome Conference, Lorne, Australia 2018 -Poster presentation.

PEDIGREE
Pathology, Epidemiology, DNA, Informatics and Genetics: A Research Enabling Enterprise
http://www.pedigree.org.au/

THE UNIVERSITY OF MELBOURNE

MONASH University

Invasive Lobular Breast Cancer : Using tumour genome-wide DNA methylation for subtyping and aid in the identification of susceptibility genes

Medha Suman[1], JiHoon Eric Joo[1,2], Ee Ming Wong[1,2], Tu Nguyen-Dumont[1,2], Neil O'Callaghan[1], Melissa Yow[1], ABCFS[3], MCCS[4], kConFab[5], John L. Hopper[3], Graham G. Giles[3, 4, 6], Roger Milne[3,4], Melissa C. Southey[1,2,4] .

[1] Genetic Epidemiology Laboratory, Department of Pathology, Victorian Comprehensive Cancer Centre, The University of Melbourne, [2] Precision Medicine, School of Clinical Sciences at Monash Health, Monash University, [3]Centre for Biostatistics and Epidemiology, School of Population and Clinical Health, The University of Melbourne, [4]Cancer Epidemiology and Intelligence Division, Cancer Council Victoria, [5]Peter MacCallum Cancer Centre, Victorian Comprehensive Cancer Centre, Melbourne, [6] School of Public Health and Preventive Medicine, Monash University

Background

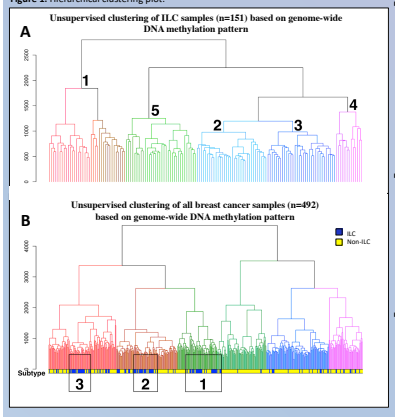
- Invasive Lobular Breast Cancer (ILC) is a distinct subtype of breast cancer and accounts for 10-15% of all breast cancer cases.
- Studies suggest that there is a strong familial risk associated with ILC, which is significantly greater than any other breast cancer subtype [1].
- A whole-genome sequencing project has been conducted in our lab involving 120 early onset and multiple-case ILC families. The extent of genetic variation in these genomes made interpretation challenging [2].
- **The aim of this project is to subtype ILC based on variation in tumour DNA methylation and to aid in the interpretation of the germline genomic data.**

Samples and Method

- The samples, lobular (n=151) and non-lobular (predominantly ductal) (n=341) in this project were sourced from :-
  - Melbourne Collaborative Cohort Study (**MCCS**)
  - Australian Breast Cancer Family Registry (**ABCFR**)
  - Kathleen Cuningham Foundation Consortium for research into Familial Breast cancer (**kConFab**)
- DNA was prepared from macrodissected formalin-fixed paraffin-embedded (FFPE) tumour sections and run on the Infinium HumanMethylation450K Beadchip array to measure methylation value at ~450,000 sites across the genome.
- All the samples were pre-processed and normalized together in R using *minfi* [3].

Results

Figure 1: Hierarchical clustering plot.



- "ward.D2" minimum variance method was used for hierarchical clustering of lobular and non-lobular breast cancer samples based on methylation pattern at 449005 CpG sites (**Fig1, A and B**).
- ER, PR and HER2 status was available for 413 out of 492 samples. The clustering was found not to be driven by the hormone receptor status.
- Lobular samples were found to cluster into three main groups (*group1*, *group2* and *group3*) in all breast cancer clustering (**Fig 1, B**).

Figure 2: Scatterplots showing difference in average methylation of lobular groups across the DMPs.

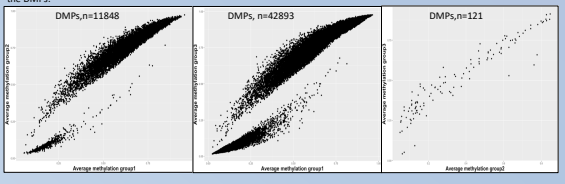
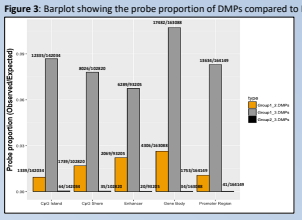
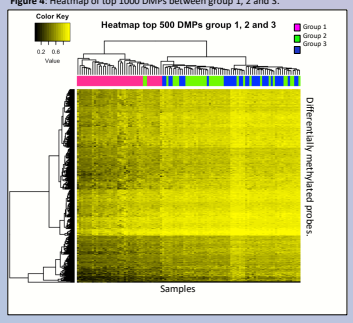


Figure 3: Barplot showing the probe proportion of DMPs compared to Illumina450K probes in different genomic locations.



- Differentially methylated probes (DMPs) were identified using *limma* [4].
- An enrichment of hypermethylated probes are present in *group 2* and *3* compared to *group1* **(Fig2)**.
- *Group 2* and *group 3* were similar in their overall methylation pattern.

Figure 4: Heatmap of top 1000 DMPs between group 1, 2 and 3.



- Differentially methylated regions (DMRs) were calculated in a pair-wise analysis between *group 1-2*, *group 1-3* and *group 2-3* using *limma*.
- We found an over representation of RNA genes differentially methylated between *group 1 and 2* (**Table 1**).
- Many of these microRNAs were found to target genes involved in important pathways related to DNA damage repair, chromatin remodelling and gene expression.

Table 1: 10 most differentially methylated regions (DMRs).

Group 1 and 2 DMRs (Total=1024)		Group 1 and 3 DMRs (Total=6435)		Group 2 and 3 DMPs (Total= 121)	
Genes and their functions	minfdr	Gene and their functions	minfdr	Gene and their functions	minfdr
MIR412, MIR369 MIR410, MIR409 MIR656, MIR541 *RNA genes*	1.14E-85	**HIC1** *Growth regulator and tumour repressor gene.*	5.88E-106	**MAFA** *Transcription factor*	2.17E-06
HTR2A *G-protein coupled receptor activity and drug binding.*	2.94E-70	**BOLL** *RNA-binding protein, expected role in germ cell development.*	1.98E-95	**EBF3** *Regulation of genes involved in cell cycle arrest and apoptosis*	3.14E-06
SMAD9 *Transcription factor activity, sequence-specific DNA binding.*	3.24E-58	**VARS** *tRNA aminoacylation and gene expression*	7.45E-91	**GSX1** *Sequence-specific DNA binding and transcriptional activator activity*	2.25E-10
MIR433, MIR127, MIR136, MIR431, RTL1, MIR432 *RNA genes.*	1.77E-51	**CD81** *Regulation of cell development, activation, growth and motility.*	6.34E-79	**NEUROG1** *Transcriptional regulator*	2.63E-09
FAM99B, AP006285.6 *RNA Genes.*	5.45E-49	**C6orf15** *Heparin binding and fibronectin binding.*	1.26E-78	**TLX3** *Transcriptional misregulation in cancer pathway*	0.001836574
MIR1182 *RNA Gene*	2.16E-48	**HTR2A** *G-protein coupled receptor activity and drug binding.*	2.94E-75	**LRFN5** *Homophilic cell-cell adhesion in a Ca(2+)-independent manner.*	0.002391575
MIR487A MIR382, MIR134, MIR323B, MIR485 *RNA Gene*	1.08E-47	**PRRC2A** *poly(A) RNA binding, may have a role in the regulation of pre-mRNA splicing.*	1.77E-74	**FAM43A** *Family With Sequence Similarity 43 Member A*	0.003113388
C6orf15 *Heparin binding and fibronectin binding.*	5.72E-47	**NOTCH4** *Evolutionarily conserved intercellular signaling pathway*	2.57E-73	**WNT2B** *Role in human development as well as carcinogenesis*	0.00446957
UBD *Proteasome binding.*	1.14E-45	**PSMB8** *Key role in apoptosis.*	3.19E-71	**PAX5** *Transcription factor activity, sequence-specific DNA binding*	0.004624716
KNDC1 *Potential role in cellular senescence.*	5.39E-44	**RGL2** *Ras signaling G-protein signalling-RAS regulation pathway.*	6.07E-68	**EN1** *Role in controlling development*	0.00481553

Summary and future directions

- Our project has identified two main subgroups of ILC based on genome-wide DNA methylation pattern.
- Combining this information with clinical outcomes, additional tumour characteristics and genomic data may help to understand ILC better and further lead to the identification of ILC susceptibility genes.

References:
1. Allen-Brady *et. al* (2005). International Journal of Cancer. 117 (4): 655–61.
2. Cancer Australia – Grant No. APP1047347 (2013-2016), *High Risk Genes for Lobular Breast Cancer.* **Southey MC**, Goldgar DE, Park DJ, Lonie A, Winship I, Simpon P, Lakhani S.
3. Aryee MJ *et. al*(2014) Bioinformatics, 30(10), pp. 1363–1369.
4. Peters TJ *et. al*(2015) Epigenetics & Chromatin, 8, pp. 6.

Additional File 4: Victorian Cancer Bioinformatics symposium: Melbourne, Australia 2019 -Poster presentation

Familial Aspects of Cancer Conference, Kingscliff, New South Wales, Australia 2019 – Poster presentation.

PEDIGREE
Pathology, Epidemiology, DNA, Informatics and Genetics: A Research Enabling Enterprise

VICTORIAN COMPREHENSIVE CANCER CENTRE | THE UNIVERSITY OF MELBOURNE | MONASH University | Cancer Council Victoria

http://www.pedigree.org.au/

Genome-wide Variably Methylated Tumour DNA Regions and Association with Overall Survival in Invasive Lobular Breast Cancer

Medha Suman[1], Pierre-Antoine Dugué[2,3,4], JiHoon Eric Joo[1], Ee Ming Wong[1,2], Catriona McLean[5], Tu Nguyen-Dumont[1,2], John L. Hopper[4], Graham G. Giles[3, 4, 6], Roger L. Milne[2,3,4], Melissa C. Southey[1,2,3].

[1] Department of Clinical Pathology, The University of Melbourne, [2]Precision Medicine, School of Clinical Sciences at Monash Health, Monash University, [3]Cancer Epidemiology Division, Cancer Council Victoria, [4]Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, [5]Department of Anatomical Pathology, Alfred Hospital, Melbourne, [6] School of Public Health and Preventive Medicine, Monash University

BACKGROUND AND AIMS

- Invasive lobular breast cancer (ILBC) is the second most common histological subtype of breast cancer accounting for 10-15% of all cases.
- Studies have shown that tumour DNA methylation signatures have potential to be used as a prognostic and diagnostic biomarker. However, very limited data exists for ILBC.
- In this study, we aimed to investigate the genome-wide variability of DNA methylation levels across ILBC tumours and to assess their association with overall survival.

METHODS

- Tumour DNA was prepared by macrodissecting formalin-fixed paraffin embedded (FFPE) tumour tissue sections for 130 ILBC cases (sourced from the Melbourne Collaborative Cohort Study (MCCS) and genome-wide methylation was measured using Illumina HumanMethylation 450K (HM450K) BeadChip array.
- Variably methylated regions (VMRs) were identified using the R package *DMRcate* [2].
- Cox proportional hazards regression models were used to assess the association between methylation levels at the 10 most significant VMRs and overall survival.
- Replication of the VMR and survival analyses findings was examined using data retrieved from The Cancer Genome Atlas (TCGA) [3] for 168 ILBC cases. We also examined the correlation between methylation and gene expression for the 10 VMRs of interest using the TCGA data.

RESULTS

- 2,771 regions were identified across the genome with variable methylation levels.
- The 10 most significant VMRs were located in the promoter regions of *ISM1*, *APC*, *TMEM101*, *ASCL2*, *HIST3H2A*, *CELF2*, *HES5*, *NKX6*, *HCG4P3* and *EFCAB4B* (Figure 1).
- In TCGA the stronger VMRs remained significant.

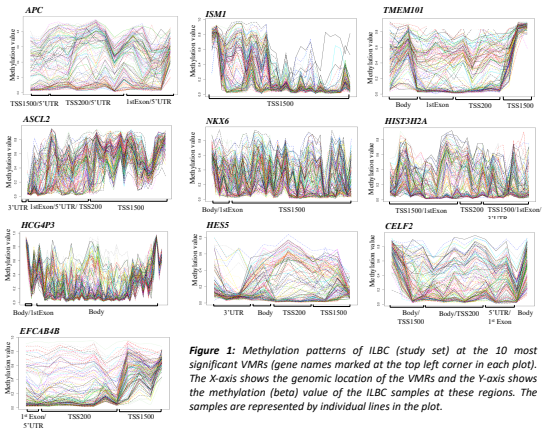


Figure 1: *Methylation patterns of ILBC (study set) at the 10 most significant VMRs (gene names marked at the top left corner in each plot). The X-axis shows the genomic location of the VMRs and the Y-axis shows the methylation (beta) value of the ILBC samples at these regions. The samples are represented by individual lines in the plot.*

RESULTS (continued)

Table 1: *Table lists the hazard ratios (HR) for the association between the methylation levels at the 10 most significant VMRs and overall survival in the Study set, its replication in the TCGA dataset and a pooled analysis of the individual studies.*

	Study set (n=130) Recorded deaths (n= 37)		TCGA dataset (n=168) Recorded deaths (n = 14)		Pooled analysis	
Gene*	HR (95% CI) (Adjusted for age and tumour stage)	*P*	HR (95% CI) (Adjusted for age and tumour stage)	*P*	HR (95% CI)	*P*
APC	1.24 (1.04- 1.49)	0.01	1.06 (0.82- 1.38)	0.63	1.18 (1.02-1.36)	0.03
TMEM101	1.22 (0.99- 1.51)	0.06	1.27 (0.87- 1.85)	0.21	1.23 (1.02-1.48)	0.03
ISM1	0.90 (0.65- 1.26)	0.54	1.48 (0.86- 2.54)	0.15	1.03 (0.77-1.36)	0.83
ASCL2	0.99 (0.71- 1.38)	0.95	1.44 (0.81- 2.57)	0.22	1.08 (0.81-1.45)	0.57
HIST3H2A	1.03 (0.82- 1.29)	0.78	1.23 (0.90- 1.68)	0.18	1.09 (0.91-1.31)	0.33
NKX6	1.01 (0.79- 1.29)	0.91	2.01 (1.28- 3.17)	0.002	1.18 (0.95-1.46)	0.13
HCG4P3	1.25 (0.91- 1.72)	0.16	1.69 (1.05- 2.72)	0.03	1.37 (1.05-1.79)	0.02
HES5	1.13 (0.89- 1.42)	0.29	1.13 (0.76- 1.68)	0.53	1.13 (0.92-1.38)	0.23
CELF2	1.13 (0.93- 1.36)	0.21	1.51 (1.07- 2.13)	0.02	1.21 (1.02-1.43)	0.02
EFCAB4B	0.99 (0.83- 1.19)	0.99	1.25 (0.93- 1.67)	0.14	1.05 (0.90-1.23)	0.49

* Genes associated with the VMRs.

- Higher methylation at *APC*, *TMEM101* and *HCG4P3* was found to be strongly associated with reduced overall survival in ILBC women in the Study set.
- In the TCGA, all the association were consistent and were in similar direction as the study set.
- Pooled analysis of the two individual studies revealed that *APC*, *TMEM101*, *HCG4P3* and *CELF2* have predictive value for patient outcome in ILBC women.
- Methylation at *ISM1*, *HIST3H2A*, *ASCL2*, *CELF2*, *EFCAB4B* and *HES5* showed a strong negative correlation with gene expression.

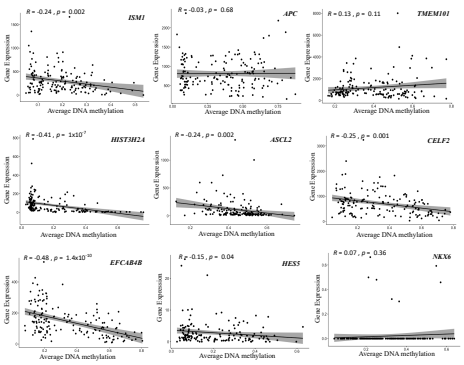


Figure 2: Correlation between average DNA methylation (shown on the x-axis) and gene expression levels (shown on the y-axis) at the genes associated with the most significant VMRs in ILBC (TCGA dataset).

CONCLUSION AND FUTURE DIRECTIONS

- This study indicates that methylation level at the variable regions may explain differences in tumour prognosis within the ILBC subtype.
- *APC*, *TMEM101*, *HCG4P3* and *CELF2* showed prognosis predictive value in ILBC.
- Further studies are needed to confirm our findings and access their utility in a clinical setting.

References :-

1. Cancer Australia – Grant No. APP1047347 (2013-2016), High Risk Genes for Lobular Breast Cancer. Southey MC, Goldgar DE, Park DJ, Lonie A, Winship I, Simpon P, Lakhani S.
2. Peters TJ, Buckley MJ, Statham AL, Pidsley R, Samaras K, Lord RV, Clark SJ, Molloy PL (2015). "De novo identification of differentially methylated regions in the human genome." *Epigenetics & Chromatin*, 8, 6
3. Cancer Genome Atlas Research Network. Weinstein JN, Collisson EA, Mills GB, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45:1113–20.

Additional File 5: Publication

Association of variably methylated tumour DNA regions with overall survival for invasive lobular breast cancer.

Suman *et al. Clin Epigenet* (2021) 13:11
https://doi.org/10.1186/s13148-020-00975-6

Clinical Epigenetics

RESEARCH Open Access

Association of variably methylated tumour DNA regions with overall survival for invasive lobular breast cancer

Medha Suman[1,2], Pierre-Antoine Dugué[2,3,4], Ee Ming Wong[1,2], JiHoon Eric Joo[1], John L. Hopper[3,4], Tu Nguyen-Dumont[1,2], Graham G. Giles[2,3,4], Roger L. Milne[2,3,4], Catriona McLean[5] and Melissa C. Southey[1,2,3*]

Abstract

Background: Tumour DNA methylation profiling has shown potential to refine disease subtyping and improve the diagnosis and prognosis prediction of breast cancer. However, limited data exist regarding invasive lobular breast cancer (ILBC). Here, we investigated the genome-wide variability of DNA methylation levels across ILBC tumours and assessed the association between methylation levels at the variably methylated regions and overall survival in women with ILBC.

Methods: Tumour-enriched DNA was prepared by macrodissecting formalin-fixed paraffin embedded (FFPE) tumour tissue from 130 ILBCs diagnosed in the participants of the Melbourne Collaborative Cohort Study (MCCS). Genome-wide tumour DNA methylation was measured using the HumanMethylation 450K (HM450K) BeadChip array. Variably methylated regions (VMRs) were identified using the *DMRcate* package in R. Cox proportional hazards regression models were used to assess the association between methylation levels at the ten most significant VMRs and overall survival. Gene set enrichment analyses were undertaken using the web-based tool *Metaspace*. Replication of the VMR and survival analysis findings was examined using data retrieved from The Cancer Genome Atlas (TCGA) for 168 ILBC cases. We also examined the correlation between methylation and gene expression for the ten VMRs of interest using TCGA data.

Results: We identified 2771 VMRs ($P < 10^{-8}$) in ILBC tumours. The ten most variably methylated clusters were predominantly located in the promoter region of the genes: *ISM1*, *APC*, *TMEM101*, *ASCL2*, *NKX6*, *HIST3H2A/HIST3H2BB*, *HCG4P3*, *HES5*, *CELF2* and *EFCAB4B*. Higher methylation level at several of these VMRs showed an association with reduced overall survival in the MCCS. In TCGA, all associations were in the same direction, however stronger than in the MCCS. The pooled analysis of the MCCS and TCGA data showed that methylation at four of the ten genes was associated with reduced overall survival, independently of age and tumour stage; *APC:* Hazard Ratio (95% Confidence interval) per one-unit *M*-value increase: 1.18 (1.02–1.36), *TMEM101*: 1.23 (1.02–1.48), *HCG4P3*: 1.37 (1.05–1.79) and *CELF2*: 1.21 (1.02–1.43). A negative correlation was observed between methylation and gene expression for *CELF2* ($R = -0.25$, $P = 0.001$), but not for *TMEM101* and *APC*.

Conclusions: Our study identified regions showing greatest variability across the ILBC tumour genome and found methylation at several genes to potentially serve as a biomarker of survival for women with ILBC.

*Correspondence: melissa.southey@monash.edu
[2] Precision Medicine, School of Clinical Sciences at Monash Health, Monash University, Clayton, VIC 3168, Australia
Full list of author information is available at the end of the article

Introduction

Invasive lobular breast cancer (ILBC) is the second most common histological subtype of breast cancer accounting for 10–15% of all cases [1–3]. ILBCs are typically

oestrogen receptor (ER) and progesterone receptor (PR) positive and human epidermal growth factor receptor 2 (HER2) negative and are strongly associated with hormonal risk factors for breast cancer [4–7]. The incidence of ILBC increased sharply in the late 1990s as a consequence of the increased use of hormone replacement therapy (HRT) [8–13]. Awareness of the increased risk of breast cancer associated with HRT led to reduced use and a decline in ILBC incidence [14], but it has been shown to increase again recently [15, 16].

ILBCs display an obscure growth pattern with small, round and discohesive cells growing in a single file without forming any distinct clusters [17]. This is likely to be related to a loss of E-cadherin protein which is common in ILBC tumourigenesis and is a hallmark of this subtype [18]. Compared with other breast cancer types, ILBCs are less likely to form a firm and distinct lump and often present as undefined palpable masses on mammography [19, 20]. This poses a significant challenge for its early detection by routine mammographic screening [19, 21–23]. This occult nature may explain the detection of ILBC cases at advanced stages [4, 24–26]. ILBCs display a unique metastatic behaviour and often metastasis to the gastrointestinal tract [27, 28], colon [29], ovaries [30] and uterus [31], which is uncommon for other breast cancers types.

ILBC is biologically and histologically heterogeneous with several histological subtypes described that show distinct clinical behaviour and outcomes [3, 17, 32–35]. Aberrant tumour DNA methylation is a hallmark of cancer that occurs early in cancer development and is thus a potentially valuable marker of tumour progression and patient survival. Alterations in tumour DNA methylation have been investigated in detail for many types of cancer, including breast cancer but ILBCs are largely underrepresented in these studies [36, 37]. Studies focusing on ILBC-specific DNA methylation alterations have mainly used a candidate gene approach and have reported aberrant promoter methylation status for specific genes such as *CDH1* [38–41], *RASSF1A, HIN-1, RAR-β, cyclin-D2, TWIST* [42], *ADAM33* [43], *SFRP1* [44] and *DAPK1* [45]. Moelans et al. (2015) compared the methylation profiles of classic ILBC ($n=20$), pleomorphic ILBC ($n=16$) and IDBC ($n=20$) for 24 established and putative tumour suppressor genes and found lower *TP73* and *MLH1* promoter methylation and higher *RASSF1* promoter methylation in pleomorphic compared with classic ILBC [46]. Bae et al. (2004) compared the methylation profiles of ILBC ($n=19$), IDBC ($n=60$) and mucinous breast cancer ($n=30$) for a panel of 12 genes and found *BRCA1* promoter hypermethylation in 92% of mucinous breast cancer compared with 39% in ILBC and 28% in IDBC. They also reported ILBC and mucinous breast cancer samples to be more frequently methylated for other genes in the panel compared with IDBC [47].

In this study, we hypothesised that genome-wide variations in DNA methylation patterns within the ILBC group may guide or reflect different tumour biologies leading to subgroups of tumours that differ in their clinical behaviour. Our aims were twofold: i) to investigate the genome-wide DNA methylation variability within the ILBC group and ii) to assess associations between tumour methylation at the most variable methylated regions and overall survival for women with ILBC.

Results

Study participants

The median age at breast cancer diagnosis in the MCCS was 65 years with tumours being diagnosed at stage 1A/1B (50%), 2A/2B (37%) and 3A/3C/4 (9%). There were 37 deaths observed during follow-up (median [IQR]: 13 [9–18] years). The tumours were mainly ER-positive, PR-positive and HER2-negative (47%). In TCGA data, the median age at diagnosis was 62 years. In both datasets, older women (aged 60 years or older at diagnosis) formed the majority of the cases (65%, in the MCCS and 58%, in TCGA). There was a higher proportion of young women at diagnosis (age less than 50 years: 21%) in TCGA compared with the MCCS (5%). The proportion of later-stage tumours (3A/3B/3C/4) was also higher in TCGA (33%) compared with the MCCS (9%). A total of 14 deaths were recorded during the follow-up (median [IQR]: 2 [1.5–5] years) in TCGA dataset. The clinical and pathological features of the study participants in the MCCS and TCGA and a comparison of the two studies are summarised in Table 1.

Variably methylated regions in ILBC

We identified 2,771 regions across the genome that showed substantially variable methylation ($P<10^{-8}$) across ILBCs in the MCCS (Additional file 2: Table S1). These VMRs corresponded to 2,208 genes and 563 intergenic regions. The most significant regions ($P<10^{-8}$) and the genes associated with these regions were chr20:13199787–13201844 (*ISM1*, 29 CpGs), chr5:112073348–112074043 (*APC*, 16 CpGs), chr17:42091713–42093050 (*TMEM101*, 16 CpGs), chr11:2290953–2293552 (*ASCL2*, 41 CpGs), chr10:134598496–134602228 (*NKX6*, 39 CpGs) and chr1:22844750–228647248 (*HIST3H2A/HIST3H2BB*, 28 CpGs). The average methylation level (beta-values) ranged between 0.09 and 0.63 at *ISM1*, 0.08 and 0.82 at *APC*, 0.15 and 0.83 at *TMEM101*, 0.15 and 0.77 at *ASCL2*, 0.07 and 0.70 at *NKX6*, and 0.05 and 0.58 at *HIST3H2A/ HIST3H2BB* (Fig. 1). There was some tendency for VMRs including more CpGs to be more highly ranked (Additional file 1: Fig. S1). We

Table 1 Clinical and pathological features of the study participants from the MCCS and TCGA

Sample characteristics	MCCS $N=130$	TCGA $N=168$	P value
Median age at diagnosis (years), interquartile range	65 [25%; 58]	62 [25%; 51]	0.02
< 50 years (*n*, %)	6 (5)	35 (21)	0.0002
50–60 years (*n*, %)	39 (30)	35 (21)	
60+ years (*n*, %)	85 (65)	98 (58)	
Year of diagnosis (*n*, %)			
1992–1996	18 (14)	0 (0)	4.4×10^{-30}
1997–2001	47 (36)	4 (2)	
2002–2005	36 (28)	15 (9)	
2006 and later	29 (22)	147 (86)	
Missing	0 (0)	2 (1)	
Overall deaths (*n*, %)	37 (28)	14 (8)	4.7×10^{-06}
Median follow-up time (years)	13	2	2.2×10^{-16}
Tumour grade (*n*, %)			
Grade I	13 (10)	NA	NA
Grade II	80 (61)	NA	
Grade III	17 (13)	NA	
Missing	20 (15)	NA	
Tumour stage (*n*, %)			
1A/1B	65 (50)	20 (12)	1.9×10^{-12}
2A/2B	48 (37)	92 (55)	
3A/3C/4	17 (13)	55 (33)	
Missing	0 (0)	1 (0.5)	
Tumour ER expression (*n*, %)			
Positive	121 (93)	157 (93)	0.32
Negative	8 (6)	6 (4)	
Missing	1 (1)	5 (3)	
Tumour PR expression (*n*, %)			
Positive	94 (72)	140 (83)	0.004
Negative	35 (27)	22 (13)	
Missing	1 (1)	6 (4)	
Tumour HER2 expression (*n*, %)			
Positive	11 (8)	21 (13)	1.5×10^{-5}
Negative	92 (71)	84 (50)	
Equivocal	5 (4)	35 (21)	
Missing	22 (17)	28 (17)	

ER oestrogen receptor, *PR* progesterone receptor, *HER2* human epidermal growth factor receptor 2

P = values are for chi-square tests and T-tests for categorical and continuous variables, respectively

found a significant enrichment for CpG island-associated regions compared to all probes included in the HM450K array (Fig. 2a). Gene annotation also showed that 62% of the VMRs were located in gene promoter regions (1st Exon, 5 prime UTR, TSS1500 and TSS200) compared with 20% in gene body regions and 23% in enhancer regions (Fig. 2b). The pathway enrichment analysis showed that the genes associated with the VMRs were enriched for 1,973 terms (FDR-adjusted $P<0.05$) including 54 KEGG pathways with stronger evidence for *neuroactive ligand-receptor interaction* (hsa04080), *breast cancer* (hsa05224), *pathways in cancer* (hsa05200), *hippo signalling pathway* (hsa04390), *Rap1 signalling pathway* (hsa04015) and *PI3K-Akt signalling pathway* (hsa04151). Figure 3 shows the twenty most significant KEGG pathways enriched in the VMRs.

Replication of the VMR analysis in TCGA dataset ($n=168$), identified 2760 VMRs, of which 763 (28%) overlapped with the MCCS. The ten most significant VMRs identified in the MCCS ranked highly in the TCGA dataset (Table 2).

Pathway enrichment analysis of the 763 overlapping VMRs resulted in 416 enriched functional terms (FDR-adjusted $P<0.05$) including nine enriched KEGG pathways. Of these, 369 overlapped with pathways identified for all MCCS VMRs; *neuroactive ligand-receptor interaction* (hsa04080) and *hippo signalling pathway* (hsa04390) were among the KEGG pathways that were also found to be significantly enriched using all MCCS VMRs.

VMRs and association with overall survival

In the MCCS, higher tumour methylation showed association with shorter overall survival for *APC* (HR = 1.28, 95% CI: 1.07–1.53), *HIST3H2A/HIST3H2BB* (HR = 1.28, 95% CI: 1.02–1.62), *CELF2* (HR = 1.30, 95% CI: 1.07–1.58) and *TMEM101* (HR = 1.21, 95% CI: 1.00–1.48). Weak evidence of association was also observed for *ISM1* (HR = 1.34, 95% CI: 0.97–1.85), *NKX6* (HR = 1.25, 95% CI: 0.98–1.60) and *HCG4P3* (HR = 1.24, 95% CI: 0.93–1.67). After adjusting for age at diagnosis and tumour stage, the association remained consistent for *APC* (HR = 1.24, 95% CI: 1.04–1.49), *TMEM101* (HR = 1.22, 95% CI: 0.99–1.51) and *HCG4P3* (HR = 1.25, 95% CI: 0.91–1.72) (Table 3). As shown in Table 3, all VMRs had an average methylation level below 0.5 and the direction of association was positive (gains in methylation associated with shorter survival).

In TCGA dataset, the crude HRs were all positive, consistent with the MCCS dataset, albeit generally greater, in particular for *ISM1* (HR = 1.48, 95% CI: 0.91–2.41), *ASCL2* (HR = 1.28, 95% CI: 0.74–2.20), *NKX6* (HR = 2.06, 95% CI: 1.32–3.21), *HIST3H2A/HIST3H2BB* (HR = 1.35, 95% CI: 1.00–1.83), *HCG4P3* (HR = 2.04, 95% CI: 1.32–3.15), *CELF2* (HR = 1.50, 95% CI: 1.06–2.12) and *EFCAB4B* (HR = 1.41, 95% CI: 1.05–1.89). Associations remained consistent after adjustment for age at diagnosis and tumour stage for all VMRs except those located

(See figure on next page.)

Fig. 1 Methylation pattern of invasive lobular breast cancer (ILBC) samples. Heatmaps show the methylation patterns of invasive lobular breast cancer (ILBC) samples in the Melbourne Collaborative Cohort Study (MCCS) across the ten most significant variably methylated regions (VMRs): **a** *ISM1*, **b** *APC*, **c** *TMEM101*, **d** *ASCL2*, **e** *NKX6*, **f** *HIST3H2A*, **g** *HCG4P3*, **h** *HES5*, **i** *CELF2* **j** *EFCAB4B*. Annotation of CpGs by genomic position and location in the context of gene are marked on the maps. Annotation of samples by age at diagnosis and tumour characteristics are shown in the colour bars as indicated in the legend on the top-right. The methylation beta-value of the CpG positions shown in the heatmap is indicated in the colour key on the top-right corner

at *APC* and *HES5*. The pooled HRs after adjustment for age at diagnosis and tumour stage showed that methylation was associated with overall survival for four genes: *APC* (HR = 1.18, 95% CI: 1.02–1.36), *TMEM101* (HR = 1.23, 95% CI: 1.02–1.48), *HCG4P3* (HR = 1.37, 95% CI: 1.05–1.79) and *CELF2* (HR = 1.21, 95% CI: 1.02–1.43) (Table 4).

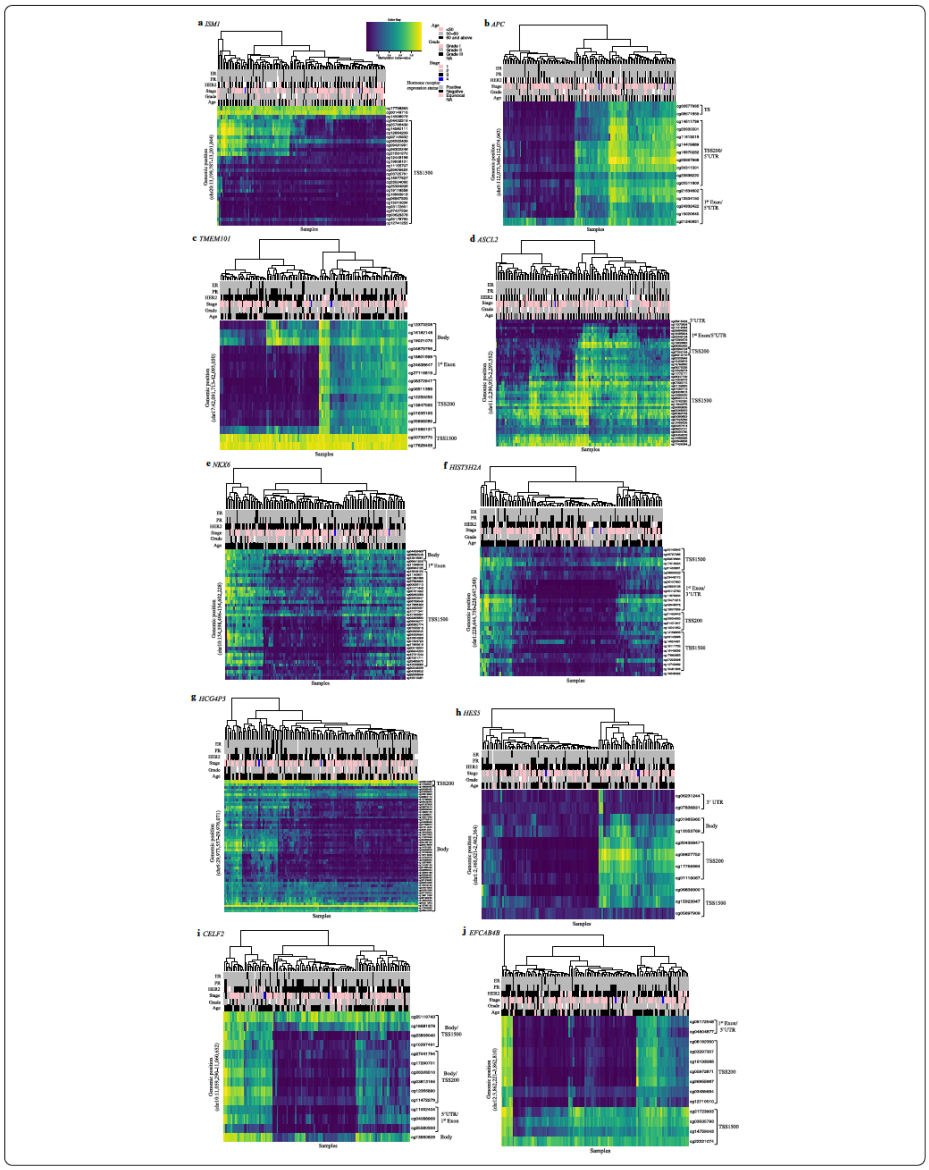
Correlation with gene expression

A relatively strong negative correlation between DNA methylation and gene expression was observed for six of the nine tested VMRs in TCGA (Fig. 4). These included *EFCAB4B* ($R = -0.5$, $P = 1.4 \times 10^{-10}$), *CELF2* ($R = -0.25$, $P = 0.001$), *HIST3H2A* ($R = -0.41$, $P = 1 \times 10^{-7}$), *ASCL2* ($R = -0.24$, $P = 0.002$), *ISM1* ($R = -0.24$, $P = 0.002$) and *HES5* ($R = -0.15$, $P = 0.04$). No or slightly positive correlation between DNA methylation and gene expression levels was observed for *APC*, *TMEM101* and *NKX6*. The feature-by-feature analysis of correlations with gene expression was very consistent with the analysis using average methylation, virtually all associations being in the same direction, with only moderate variation in effect estimates (Additional file 3: Table S2).

Discussion

We investigated the genome-wide DNA methylation pattern of ILBC tumours, with the aim of identifying methylation markers predictive of patient outcome. Scanning of the ILBC methylome revealed regions of variable methylation in ILBC tumours. The VMRs were primarily located in CpG island regions and were significantly enriched in pathways such as *breast cancer* (hsa05224), *pathways in cancer* (hsa05200), *hippo signalling pathway* (hsa04390), *Rap1 signalling pathway* (hsa04015) and *PI3K-Akt signalling pathway* (hsa04151). These pathways have previously been found to be dysregulated in cancer tissue [48–53]. Some of the key genes involved in the enriched pathways included *APC, DAPK1, BMP2 and CCND2. DAPK1* is an important regulator of cell apoptotic pathways [54] and *DAPK1* promoter hypermethylation has previously been reported in ILBCs with a potential role in tumour progression [45, 55]. *BMP2* is a member of the TGF-ß superfamily and is involved in cell proliferation and differentiation during tumour formation [56]. Promoter methylation of *BMP2* has been associated with breast cancer progression and drug resistance [57]. *CCND2* promoter methylation was previously reported to be a common event in breast cancer and have prognostic value [58]. We found a similar DNA methylation variability profile in TCGA dataset, in particular for the VMRs showing strongest variability in the MCCS.

Several previous studies have reported tumour DNA methylation to have prognostic value in cancer [59–64]. Methylation at many gene promoters has been reported to have independent prognostic value in breast cancer including *HOXA11* [65], *ESR1* and *PITX2* [66], *HOXD13* [67] *CDH22* [68] *BRCA1* and *RASSF1* [69, 70]. Tumour DNA methylation and its prognostic significance has also been investigated for certain breast cancer subtypes, in particular gene expression-based subtypes. Thomas et al. (2017) used hierarchical clustering based on DNA methylation to further segregate luminal A tumours into two subgroups and found that the subgroup with lower relative methylation showed better prognosis [71], similar to the findings of our study. Another study using whole-genome methylation sequencing stratified triple-negative breast cancers into three methylation-defined clusters and found the hypomethylated cluster to show better prognosis compared with the other two highly methylated clusters [72], also consistent with our results. However, to our knowledge, no study has reported on the overall tumour methylation variability in ILBC and tested the potential for the variably methylated regions to be used as prognostic markers. The assessment of VMRs was genome-scale but only the highest ranking VMRs were tested for their association with survival. Although many of the tested VMRs showed a significant association with overall survival, there could be other VMRs or individual CpG sites for which methylation is associated with survival. We found that promoter hypermethylation at *APC*, *TMEM101* and *HCG4P3* was associated with shorter overall survival in the MCCS after adjustment for age and tumour stage. The results in TCGA were largely consistent with the MCCS, although associations generally appeared stronger; this might suggest that the prognostic value of these DNA methylation markers is greater for women with more advanced ILBC. In the pooled analysis, DNA methylation at four genes (*APC, TME101,*


a ISM1
b APC
c TMEM101
d ASCL2
e NKX6
f HIST1H2A
g HCG4P3
h HES5
i CELF2
j EPCAM-AS1

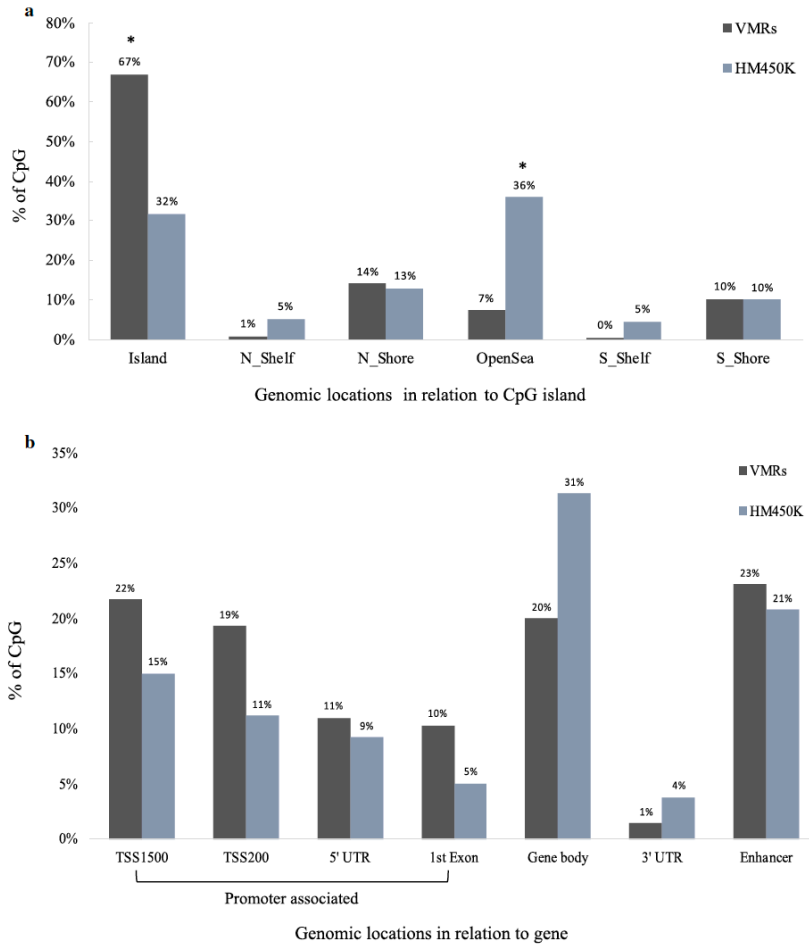


Fig. 2 Genomic distribution of the variably methylated regions (VMRs). Bar plots show the distribution of 2771 variably methylated regions (VMRs) identified within invasive lobular breast cancer (ILBC) samples in the Melbourne Collaborative Cohort Study (MCCS) **a** relative to CpG islands, shores (0–2 kb from island), shelves (2–4 kb from island) and open sea and **b** in relation to the gene. Different genomic locations are shown on the x-axis and the percentage of CpG positions related to the VMRs is shown on the y-axis. The distribution of the HM450K probes relative to each CpG context is also indicated. *P*-values (Chi-square test) assessing significant enrichment in a given category relative to the HM450K array composition are indicated (*$P < 0.001$)

HCG4P3 and *CELF2*) was associated with shorter overall survival. All the highest-ranking VMRs had an average methylation level below 0.5, and the direction of association with survival was virtually always positive, which indicates that methylation gains (i.e. loss of the normal hypomethylation state) were associated with worse survival. *APC* is a well-known tumour suppressor gene and this finding is in agreement with previous reports [73,

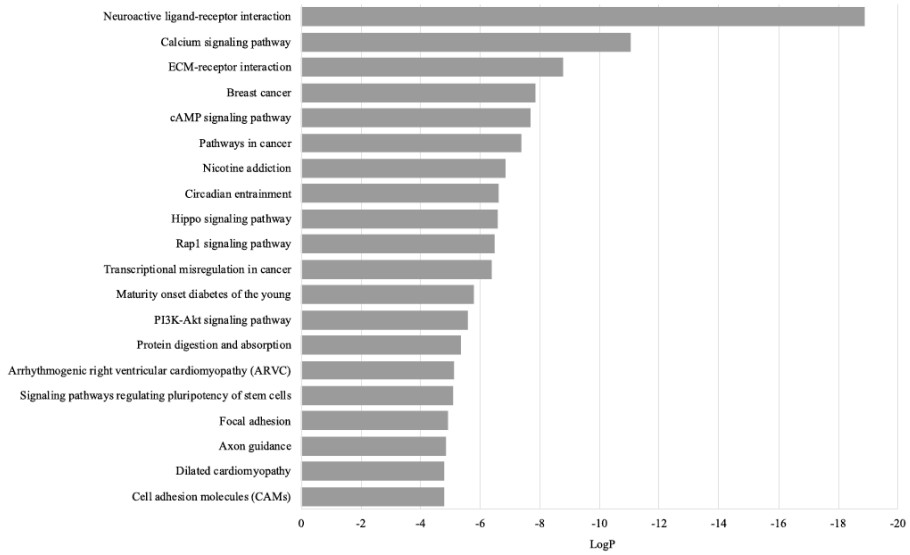


Fig. 3 Twenty most significantly enriched KEGG pathways. Bar plot shows twenty most significantly enriched KEGG pathways in the variably methylated region (VMRs) identified within invasive lobular breast cancer (ILBC) samples in the Melbourne Collaborative Cohort Study (MCCS). The enriched terms are shown on the *y*-axis and the *P*-values (log transformed) assessing significant enrichment are shown on the *x*-axis

Table 2 Ten most significant VMRs identified in the MCCS and their respective ranking in TCGA

MCCS				TCGA		
Genomic location of the VMRs (GRCh37)	minfdr*	Number of CpGs	Associated gene	Genomic location of the VMRs in relation to the corresponding genes	minfdr*	Rank in TCGA
chr20:13199787–13201844	5×10^{-181}	29	*ISM1*	TSS1500	1×10^{-120}	10
chr5:112073348–112074043	5×10^{-181}	16	*APC*	Body, 1st exon, TSS200, TSS1500	3×10^{-170}	4
chr17:42091713–42093050	4×10^{-172}	16	*TMEM101*	TSS1500, TSS200, 5′UTR, 1st exon	3×10^{-92}	20
chr11:2290953–2293552	2×10^{-152}	41	*ASCL2*	3′UTR, 1st exon, 5′UTR, TSS200, TSS1500	1×10^{-90}	23
chr10:134598496–134602228	1×10^{-142}	39	*NKX6*	Body, 1stExon, TSS1500	6×10^{-118}	12
chr1:228644750–228647248	1×10^{-131}	28	*HIST3H2A/HIST3H2BB*	TSS1500, TSS200	2×10^{-196}	2
chr6:29973557–29976071	4×10^{-124}	52	*HCG4P3/HLA-J*	Body	2×10^{-194}	3
chr1:2460621–2462364	1×10^{-110}	11	*HES5*	3′UTR, Body, TSS200, TSS1500	3×10^{-90}	24
chr10:11059290–11060652	2×10^{-109}	14	*CELF2*	TSS1500, TSS200, 5′UTR, 1st exon	9×10^{-130}	7
chr12:3862221–3862810	6×10^{-104}	13	*EFCAB4B*	1stExon, 5′UTR, TSS200, TSS1500	7×10^{-198}	1

*minfdr: minimum adjusted *P*-value, TSS200 is the region from Transcript start site (TSS) to 200 nucleotides (nt) upstream of TSS; TSS1500 is the region from 200 to 1500 nt upstream of TSS; 5′ UTR is the region within 5 prime untranslated region, between the TSS and the ATG start site; body is the region between the ATG and stop codon; 3′ UTR is between the stop codon and poly A signal

74]. Debouki et al. (2017) found a significant correlation between *APC* promoter methylation and aggressive behaviour of both non-familial and familial breast cancer in the Tunisian population [73]. The association of *APC* promoter methylation with reduced survival has also been reported for other cancer types, such as

Table 3 Hazard ratios (HRs) for the association between the methylation levels at the ten most significant variably methylated regions (VMRs) and overall survival in the Melbourne Collaborative Cohort Study (MCCS) and The Cancer Genome Atlas (TCGA) dataset

MCCS							TCGA						
				Adjusted for age		Adjusted for age and stage				Adjusted for age		Adjusted for age and stage	
Gene*	Averagemethylation**	HR (95% CI)	*P*	HR (95% CI)	*P*	HR (95% CI)	*P*	Average methylation**	HR (95% CI)	*P*	HR (95% CI)	*P*	HR (95% CI)	*P*
APC	39.1	1.28 (1.07–1.53)	0.01	1.24 (1.04–1.47)	0.02	1.24 (1.04–1.49)	0.01	35.9	1.16 (0.89–1.51)	0.28	1.12 (0.86–1.44)	0.41	1.06 (0.82–1.38)	0.63
TMEM101	39.6	1.21 (1.00–1.48)	0.06	1.19 (0.98–1.44)	0.08	1.22 (0.99–1.51)	0.06	39	1.13 (0.77–1.66)	0.52	1.12 (0.78–1.60)	0.54	1.27 (0.87–1.85)	0.21
ISM1	22.8	1.34 (0.97–1.85)	0.07	1.02 (0.76–1.38)	0.89	0.90 (0.65–1.26)	0.54	19.2	1.48 (0.91–2.41)	0.11	1.38 (0.80–2.37)	0.25	1.48 (0.86–2.54)	0.15
ASCL2	44.4	1.16 (0.81–1.65)	0.42	0.93 (0.66–1.31)	0.69	0.99 (0.71–1.38)	0.95	44.6	1.28 (0.74–2.20)	0.38	1.17 (0.68–2.02)	0.57	1.44 (0.81–2.57)	0.22
HIST3H2A	20.1	1.28 (1.02–1.62)	0.03	1.08 (0.86–1.35)	0.53	1.03 (0.82–1.29)	0.78	20.1	1.35 (1.00–1.83)	0.05	1.28 (0.94–1.73)	0.12	1.23 (0.90–1.68)	0.18
NKX6	29	1.25 (0.98–1.60)	0.07	1.06 (0.83–1.35)	0.63	1.01 (0.79–1.29)	0.91	30.2	2.06 (1.32–3.21)	0.001	1.88 (1.21–2.92)	0.01	2.01 (1.28–3.17)	0.002
HCG4P3	29.1	1.24 (0.93–1.67)	0.14	1.13 (0.84–1.53)	0.41	1.25 (0.91–1.72)	0.16	31.1	2.04 (1.32–3.15)	0.001	1.80 (1.13–2.85)	0.01	1.69 (1.05–2.72)	0.03
HES5	19	1.11 (0.88–1.40)	0.38	1.11 (0.88–1.40)	0.37	1.13 (0.89–1.42)	0.29	20.9	1.26 (0.88–1.80)	0.21	1.15 (0.80–1.65)	0.45	1.13 (0.76–1.68)	0.53
CELF2	33.4	1.30 (1.07–1.58)	0.01	1.12 (0.93–1.36)	0.23	1.13 (0.93–1.36)	0.21	35	1.50 (1.06–2.12)	0.02	1.44 (1.01–2.05)	0.04	1.51 (1.07–2.13)	0.02
EFCAB4B	32.8	1.01 (0.83–1.23)	0.88	0.96 (0.80–1.15)	0.63	0.99 (0.83–1.19)	0.99	34.5	1.41 (1.05–1.89)	0.02	1.32 (0.98–1.78)	0.07	1.25 (0.93–1.67)	0.14

*Gene: Gene associated with the variably methylated regions (VMRs), most of the VMRs were located in the promoter region of the genes, **Average methylation: Average methylation level (beta-value) of the samples across the VMRs, *HR* hazard ratio, *CI* confidence interval

Table 4 Pooled hazard ratios for the association between methylation levels at the ten most VMRs and overall survival: meta-analysis of the MCCS and TCGA results

			Adjusted for age		Adjusted for age and stage	
Gene*	HR (95% CI)	*P*	HR (95% CI)	*P*	HR (95% CI)	*P*
APC	1.24 (1.07–1.44)	0.004	1.20 (1.04–1.39)	0.01	1.18 (1.02–1.36)	0.03
TMEM101	1.19 (1.00–1.42)	0.05	1.17 (0.99–1.39)	0.06	1.23 (1.02–1.48)	0.03
ISM1	1.38 (1.05–1.80)	0.02	1.09 (0.84–1.42)	0.50	1.03 (0.77–1.36)	0.83
ASCL2	1.19 (0.88–1.61)	0.24	0.99 (0.74–1.33)	0.96	1.08 (0.81–1.45)	0.57
HIST3H2A	1.30 (1.08–1.57)	0.004	1.15 (0.95–1.37)	0.14	1.09 (0.91–1.31)	0.33
NKX6	1.40 (1.13–1.74)	0.002	1.21 (0.98–1.50)	0.08	1.18 (0.95–1.46)	0.13
HCG4P3	1.45 (1.13–1.85)	0.003	1.30 (1.00–1.67)	0.04	1.37 (1.05–1.79)	0.02
HES5	1.15 (0.95–1.40)	0.15	1.12 (0.92–1.36)	0.25	1.13 (0.92–1.38)	0.23
CELF2	1.34 (1.13–1.60)	0.0006	1.18 (1.00–1.40)	0.05	1.21 (1.02–1.43)	0.02
EFCAB4B	1.12 (0.95–1.32)	0.17	1.05 (0.89–1.22)	0.57	1.05 (0.90–1.23)	0.49

*Gene: Gene associated with the variably methylated regions (VMRs), most of the VMRs were located in the promoter region of the genes, *HR* hazard ratio, *CI* confidence interval

non-small cell lung cancer [75] and prostate cancer [76, 77]. CELF2, an RNA binding protein involved in alternative splicing, has also been reported to be involved in breast cancer growth and progression. Piqué et al., (2019) found that *CELF2* promoter methylation led to a loss of *CELF2* expression that had a growth promoter effect in breast tumours. They also found that *CELF2* promoter methylation was associated with worse patient outcome [78]. In TCGA data, we found a strong, negative correlation between *CELF2* promoter methylation and the gene expression levels. TMEM101 is a transmembrane protein that has been shown to activate NF-kappa-beta signalling pathways. There is to our knowledge no previous literature suggesting a role of *TMEM101* promoter methylation in relation to cancer progression/survival. *HCG4P3* is also known as HLA complex group 4 pseudogene 3, and there is to our knowledge no record of this gene being involved in cancer.

The main limitation of this study was the relatively small sample size that limited our analysis to all-cause death as an endpoint. The MCCS and TCGA data had different characteristics in terms of their study design and sample variation. The two studies had different follow-up times, and TCGA data had more young women and generally higher tumour stage (Table 1). Our findings for both the VMR and survival analysis were nevertheless consistent across the two studies. We considered the main factors that we thought could impact methylation profiles in tumours and ILBC survival, i.e. age and stage. Factors such as smoking, alcohol consumption or diabetes, and perhaps family history (via underlying genetic sequence) likely play some role, but it is presumably less important, so we did not include them in the analysis. These variables are not systematically collected with precision (questionnaires) in the clinical setting. In this context, our study identified methylation biomarkers, and it is likely that many factors worthy of investigation (genetic and lifestyle and environmental) play a role in explaining the observed associations. Finally, while we identified a large number of regions across the ILBC genome that showed substantial variable methylation pattern, only the strongest ten VMRs were tested for association with survival to minimise the multiple testing burden. If replicated by other studies, the methylation markers identified in our study may contribute to the development of molecular signatures for enhanced prediction of ILBC survival.

Conclusions

Our study indicates that methylation levels at the most variable regions across the genome may explain differences in tumour prognosis within the ILBC subtype. We identified *APC*, *TMEM101*, *HCG4P3* and *CELF2* promoter methylation as possibly relevant prognostic biomarkers for women with ILBC. Further studies are required to confirm our findings and to assess their utility in a clinical setting.

Methods

Study samples

The samples included in this study were obtained from the Melbourne Collaborative Cohort Study (MCCS) [79]. The MCCS was set up in 1990 with the aim of investigating the role of diet and lifestyle in cancer and other diseases. Between 1990 and 1994, 41,513 participants, aged 40–69, were recruited in the study, and baseline information on lifestyle, health and diet was collected through interviews. Women with ILBC included in this

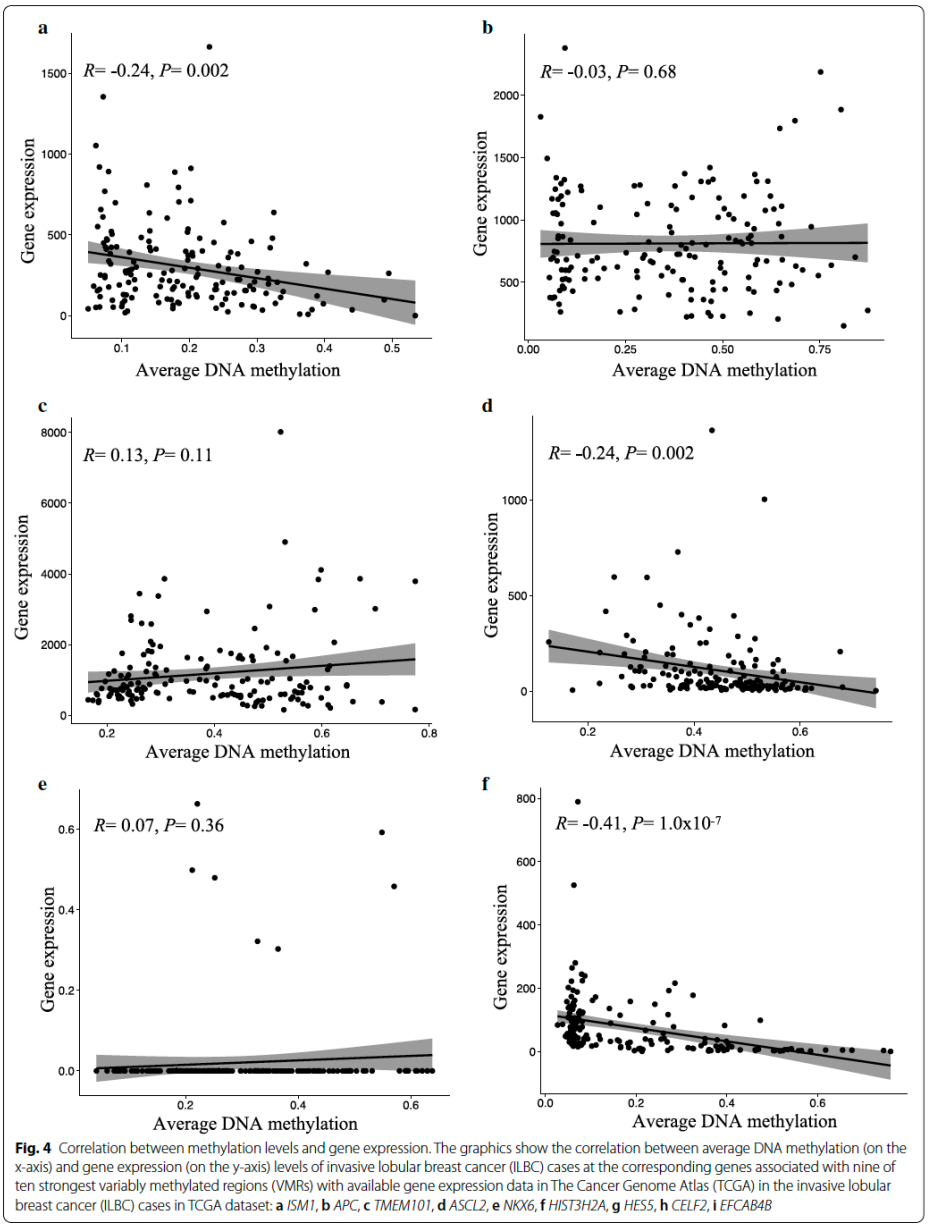


Fig. 4 Correlation between methylation levels and gene expression. The graphics show the correlation between average DNA methylation (on the x-axis) and gene expression (on the y-axis) levels of invasive lobular breast cancer (ILBC) cases at the corresponding genes associated with nine of ten strongest variably methylated regions (VMRs) with available gene expression data in The Cancer Genome Atlas (TCGA) in the invasive lobular breast cancer (ILBC) cases in TCGA dataset: **a** *ISM1*, **b** *APC*, **c** *TMEM101*, **d** *ASCL2*, **e** *NKX6*, **f** *HIST3H2A*, **g** *HES5*, **h** *CELF2*, **i** *EFCAB4B*

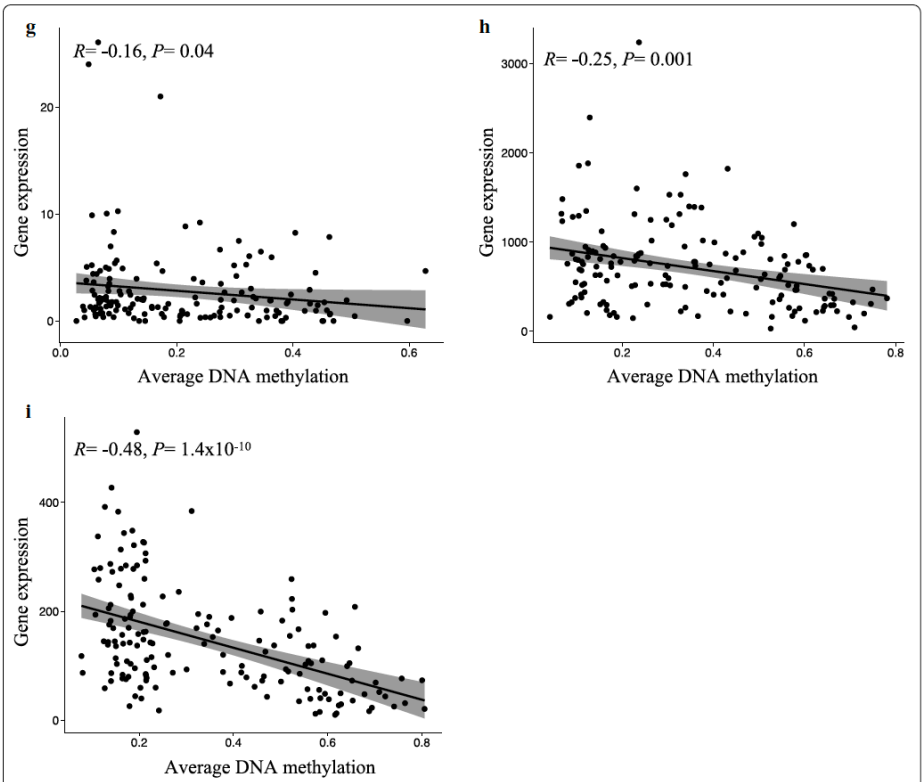


Fig. 4 continued

study were diagnosed between 1993 and 2011, based on the International Classification of Diseases for Oncology (ICD-O) codes 8520 (73%), 8522 (26%) and 8500 (1 case). Clinical and pathological characteristics of the study participants are listed in Table 1.

Endpoints

Incidences of cancer cases and deaths in the MCCS participants are regularly updated by linkage to the Victorian and national cancer and death registries, which are considered to be virtually complete. The latest linkage was completed on 31 March 2017 and death data were considered to be complete up to 31 December 2016. Overall survival was defined as the time (in years) from breast cancer diagnosis to death (from any cause) or end of follow-up.

DNA extraction from formalin-fixed paraffin embedded breast tumour tissue

Pathology material related to each ILBC case had previously been retrieved from the diagnostic service laboratory and reviewed by qualified pathologists. Unstained sections had been prepared and stored desiccated at 4 °C for up to 20 years. DNA extraction was conducted as described in Wong et al. [80]. Briefly, the tumour areas most suitable for macrodissection were identified

by a qualified pathologist using the WHO classification of tumours of the Breast Criteria (WHO Classification of Tumours of the Breast (2012). 4th edn. International Agency for Research on Cancer (IARC), Lyon) [81] and recorded by directly marking up representative H&E stained sections. An average of two corresponding 3um methyl green stained FFPE sections were macro-dissected as described in Wong et al. [82], and DNA was extracted using the QIAamp DNA FFPE protocol.

Tumour purity was estimated using the *R* tool *InfiniumPurify* [83] that takes methylation beta-values of the tumour samples and uses the methylation levels of pre-selected informative differentially methylated CpG sites (iDMCs) identified from TCGA data (when data from normal-adjacent tissue are not available) to estimate tumour purity for each tumour sample by density evaluation of Gaussian kernel. Tumour purity estimates, obtained as the proportion of tumour cells in each sample, were high, ranging from 37 to 88% across samples; 88% of the samples had an estimated tumour purity greater than 50%.

Genome-wide DNA methylation profiling

Genome-wide DNA methylation was measured using the HumanMethylation450K (HM450K) BeadChip array (Illumina). For each sample, a total of 300–500 ng of tumour DNA was bisulfite-converted using Zymo Gold EZ-DNA kit (Irvine, CA) and restored using the DNA Restoration Kit as per the manufacturer's instructions (Illumina, CA, United States). Sample DNA quantity was assessed using an in-house modified quality control protocol [80]. Samples that passed the final quality check were run on the HM450K array (Illumina) according to manufacturer's instructions.

Data pre-processing and normalisation

Raw intensity files (IDAT files) were imported into the R computing environment using the Bioconductor package *minfi* [84], and all samples were pre-processed and normalised together. Data quality was first evaluated by assessing the detection *P*-value, which was obtained for every CpG site in every sample. Samples with an average detection *P*-value > 0.01 were considered poor quality and were removed from further analysis. CpG probes with a detection *P*-value > 0.05 in at least one sample were considered unreliable and were removed from further analysis. Data were normalised using the *minfi* functional normalisation (FNORM) method to correct for both within-array (technical bias between type I and type II probes) as well as between-array unwanted variations [85]. After data pre-processing and normalisation, a total of 449,005 CpG sites remained for analysis. Beta-values (ratio of the methylated probe intensity and the sum of methylated and unmethylated probe intensity) and *M*-values ($\log_2$ beta-value) were calculated. *M*-values were used in all statistical analyses, while beta-values were used for data exploration and visualisation, as suggested in [86].

TCGA data

Raw DNA methylation data (IDAT files) for 168 ILBC cases were downloaded from the TCGA legacy database (Study Accession: phs000178) using the R package *TCGABiolinks* [87]. Methylation data were pre-processed and normalised similarly to the MCCS, and methylation values (beta-values and *M*-values) were calculated for 168 ILBCs at 440,380 CpG positions across the genome. Survival data were retrieved for 159 (95%) ILBC cases. Gene expression data in the form of normalised counts (RNA sequencing-Illumina Hi-Seq) were retrieved for 159 (95%) ILBC cases. Cases of ILBC in the TCGA dataset were diagnosed between 1992 and 2013. Clinical characteristics of the TCGA samples are listed in Table 1.

Statistical analysis

Variable methylation analysis

Variable methylation analysis was performed using the *DMRcate* package in R [88]. To identify the variably methylated regions (VMRs), the variance of *M*-values was computed across 130 ILBCs in the MCCS, and Gaussian smoothing was applied to the resulting per-CpG-site test statistics using the default *DMRcate* options. *DMRcate* uses the method of Satterthwaite to smooth test statistics and derive respective *P*-values. Nearby significant CpG sites were collapsed in clusters using a bandwidth of 1000 base pairs (bp). The clusters that showed the highest variability in DNA methylation (i.e. regions with a minimum adjusted *P*-value (min-fdr) of less than 10^{-8}) were defined as the VMRs. There were 396 solo CpGs that were not included in the VMR calculation. This analysis was replicated using the 168 ILBC samples from TCGA.

Gene set enrichment analysis

Gene set enrichment analysis was performed on all the genes associated with the VMRs using the web-based tool *Metaspace* using the default settings [89]. Pathway and gene set enrichment analysis were carried out using the KEGG Pathway database [90]. All genes in the human genome were used as the enrichment background. Pathways and biological terms with a *P*-value < 0.01,

a minimum count of 3 and an enrichment factor > 1.5 (the ratio between the observed counts and the counts expected by chance) were selected and grouped into clusters.

Survival analysis

Survival analyses were undertaken for the ten most variably methylated regions identified across the MCCS ILBC samples. Follow-up started at the date of diagnosis and ended at the date of death or end of follow-up, whichever came first. Cox proportional hazards regression models were used to calculate hazard ratios (HRs) and 95% confidence intervals (CI) for the association between DNA methylation levels (*M*-values) and risk of death. Three models were fitted: (1) univariable, with DNA methylation as a crude predictor, and multivariable, (2) with additional adjustment for age at diagnosis and (3) with adjustment for age at diagnosis and tumour stage. For each VMR, the methylation level was defined as the average methylation value across all CpG sites covering the VMR. The same analysis was carried out using the 168 ILBC samples from TCGA. Survival analyses were undertaken using the R package *Survival* [91]. HRs from the two individual studies were then pooled using fixed-effects meta-analysis with inverse variance weights.

Association with gene expression

To test if DNA methylation correlated with gene expression at the ten strongest VMRs (identified in the MCCS), we assessed the correlation between average methylation levels (average *M*-values for all CpGs covering a VMR) and gene expression levels using Pearson's correlation;we used matching gene expression and DNA methylation data available in the TCGA dataset for nine of the ten strongest VMRs.The correlations with gene expression were also assessed for individual CpG sites of each VMR.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13148-020-00975-6.

Additional file 1: Figure S1. Relation between the number of CpGs related to each VMR and the VMR ranking

Additional file 2. List of variable methylated regions identified in ILBC.

Additional file 3. Correlation between DNA methylation and gene expression (feature-by-feature analysis).

Abbreviations

CI: Confidence interval; ER: Oestrogen receptor; FFPE: Formalin fixed paraffin embedded; FNORM: Functional normalisation; HM450K: HumanMethylation 450K; HER2: Human epidermal growth factor receptor 2; HRT: Hormone replacement therapy; HR: Hazard ratio; ILBC: Invasive lobular breast cancer; ICD-O: International Classification of Diseases for Oncology; IDBC: Invasive ductal breast cancer; MCCS: Melbourne Collaborative Cohort Study; MRI: Magnetic resonance imaging; PR: Progesterone receptor; TCGA: The Cancer Genome Atlas; VMR: Variably methylated region.

Acknowledgements

The authors thank the staff of the Presidion Medicine Biorepository, School of Clinical Sciences at Monash Health, Monash University and all the participants and contributors to the Melbourne Collaborative Cohort Study.

Authors' contributions

MS generated data, conducted analyses and drafted the manuscript; P-AD led the statistical analyses and contributed to the writing of the manuscript; EMW generated data and contributed to manuscript preparation; JHJ generated data, conducted analyses, and contributed to the writing of the manuscript; JLH contributed to the preparation of the manuscript; TN-D conducted analyses and contributed to the drafting of the manuscript; GGG contributed data and contributed to the preparation of the manuscript; RLM contributed data and contributed to the preparation of the manuscript; CMcL generated data and contributed to the drafting of the manuscript; MCS conceived the study, supervised the laboratory work, conducted analyses and contributed to drafting the manuscript; All authors read and approved the final version of the manuscript.

Funding

MS was the recipient of a Beaney Scholarship in Pathology from The University of Melbourne (2018). TN-D a National Breast Cancer Foundation (Australia) Career Development Fellow (ECF-17-001). MCS is a National Health and Medical Research Council (NMHRC, Australia) Senior Research Fellow (APP1155163). This work was supported by an NHMRC Program Grant (APP1074383) and a project grant from Cancer Australia (APP1047347). Melbourne Collaborative Cohort Study (MCCS) cohort recruitment was funded by VicHealth and Cancer Council Victoria. The MCCS was further augmented by Australian National Health and Medical Research Council Grants 209057, 396414 and 1074383 and by infrastructure provided by Cancer Council Victoria. Cases and their vital status were ascertained through the Victorian Cancer Registry and the Australian Institute of Health and Welfare, including the National Death Index and the Australian Cancer Database.

Availability of data and materials

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Ethics approval and consent to participate

The participants of the Melbourne Collaborative Cohort Study provided informed consent. This study was approved by the Cancer Council Victoria Human Research Ethics Committee.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1] Department of Clinical Pathology, Melbourne Medical School, The University of Melbourne, Melbourne, VIC 3010, Australia. [2] Precision Medicine, School of Clinical Sciences at Monash Health, Monash University, Clayton, VIC 3168, Australia. [3] Cancer Epidemiology Division, Cancer Council Victoria, Melbourne, VIC 3004, Australia. [4] Centre for Epidemiology and Biostatistics, School of Population and Global Health, The University of Melbourne, Melbourne, VIC 3010, Australia. [5] Anatomical Pathology, Alfred Health, The Alfred Hospital, Melbourne, VIC 3181, Australia.

Received: 22 June 2020 Accepted: 10 November 2020
Published online: 18 January 2021

References

1. Reed AEM, McCart Reed AE, Kutasovic JR, Lakhani SR, Simpson PT. Invasive lobular carcinoma of the breast: morphology, biomarkers and 'omics. Breast Cancer Res. 2015;17(1):12.
2. Lee J-H, Park S, Park HS, Park B-W. Clinicopathological features of infiltrating lobular carcinomas comparing with infiltrating ductal carcinomas: a case control study. World J Surg Oncol. 2010;8:34.
3. Orvieto E, Maiorano E, Bottiglieri L, Maisonneuve P, Rotmensz N, Galimberti V, et al. Clinicopathologic characteristics of invasive lobular carcinoma of the breast: results of an analysis of 530 cases from a single institution. Cancer. 2008;113(7):1511–20.
4. Arpino G, Bardou VJ, Clark GM, Elledge RM. Infiltrating lobular carcinoma of the breast: tumor characteristics and clinical outcome. Breast Cancer Res. 2004;6(3):R149–56.
5. Mersin H, Yildirim E, Gülben K, Berberoğlu U. Is invasive lobular carcinoma different from invasive ductal carcinoma? Eur J Surg Oncol. 2003;29(4):390–5.
6. Bakken K, Fournier A, Lund E, Waaseth M, Dumeaux V, Clavel-Chapelon F, et al. Menopausal hormone therapy and breast cancer risk: impact of different treatments. The European Prospective Investigation into Cancer and Nutrition. Int J Cancer. 2011;128(1):144–56.
7. Flesch-Janys D, Slanger T, Mutschelknauss E, Kropp S, Obi N, Vettorazzi E, et al. Risk of different histological types of postmenopausal breast cancer by type and regimen of menopausal hormone therapy. Int J Cancer. 2008;123(4):933–41.
8. Li CI, Anderson BO, Porter P, Holt SK, Daling JR, Moe RE. Changing incidence rate of invasive lobular breast carcinoma among older women. Cancer. 2000;88(11):2561–9.
9. Verkooijen HM, Fioretta G, Vlastos G, Morabia A, Schubert H, Sappino A-P, et al. Important increase of invasive lobular breast cancer incidence in Geneva, Switzerland. Int J Cancer. 2003;104(6):778–81.
10. Li CI, Daling JR, Malone KE. Age-specific incidence rates of in situ breast carcinomas by histologic type, 1980 to 2001. Cancer Epidemiol Biomark Prev. 2005;14(4):1008–11.
11. Portschy PR, Marmor S, Nzara R, Virnig BA, Tuttle TM. Trends in incidence and management of lobular carcinoma in situ: a population-based analysis. Ann Surg Oncol. 2013;20(10):3240–6.
12. Louwman MWJ, Vriezen M, van Beek MWPM, Nolthenius-Puylaert MCB-JET, van der Sangen MJC, Roumen RM, et al. Uncommon breast tumors in perspective: incidence, treatment and survival in the Netherlands. Int J Cancer. 2007;121(1):127–35.
13. Rosenberg LU, Magnusson C, Lindström E, Wedrén S, Hall P, Dickman PW. Menopausal hormone therapy and other breast cancer risk factors in relation to the risk of different histological subtypes of breast cancer: a case–control study. Breast Cancer Res. 2006;8(1):R11.
14. Li CI, Daling JR. Changes in breast cancer incidence rates in the United States by histologic subtype and race/ethnicity, 1995 to 2004. Cancer Epidemiol Biomark Prev. 2007;16(12):2773–80.
15. Wachtel MS, Yang S, Dissanaike S, Margenthaler JA. Hormone replacement therapy, likely neither Angel Nor Demon. PLoS ONE. 2015;10(9):e0138556.
16. Ward EM, DeSantis CE, Lin CC, Kramer JL, Jemal A, Kohler B, et al. Cancer statistics: breast cancer in situ. CA Cancer J Clin. 2015;65(6):481–95.
17. Martinez V, Azzopardi JG. Invasive lobular carcinoma of the breast: incidence and variants. Histopathology. 1979;3(6):467–88.
18. Ciriello G, Gatza ML, Beck AH, Wilkerson MD, Rhie SK, Pastore A, et al. Comprehensive molecular portraits of invasive lobular breast cancer. Cell. 2015;163(2):506–19.
19. Le Gal M, Ollivier L, Asselain B, Meunier M, Laurent M, Vielh P, et al. Mammographic features of 455 invasive lobular carcinomas. Radiology. 1992;185(3):705–8.
20. Lopez JK, Bassett LW. Invasive lobular carcinoma of the breast: spectrum of mammographic, US, and MR imaging findings. Radiographics. 2009;29(1):165–76.
21. Krecke KN, Gisvold JJ. Invasive lobular carcinoma of the breast: mammographic findings and extent of disease at diagnosis in 184 patients. AJR Am J Roentgenol. 1993;161(5):957–60.
22. Berg WA, Gutierrez L, NessAiver MS, Carter WB, Bhargavan M, Lewis RS, et al. Diagnostic accuracy of mammography, clinical examination, US, and MR imaging in preoperative assessment of breast cancer. Radiology. 2004;233(3):830–49.
23. Munot K, Dall B, Achuthan R, Parkin G, Lane S, Horgan K. Role of magnetic resonance imaging in the diagnosis and single-stage surgical resection of invasive lobular carcinoma of the breast. Br J Surg. 2002;89(10):1296–301.
24. Cristofanilli M, Gonzalez-Angulo A, Sneige N, Kau S-W, Broglio K, Theriault RL, et al. Invasive lobular carcinoma classic type: response to primary chemotherapy and survival outcomes. J Clin Oncol. 2005;23(1):41–8.
25. Molland JG, Donnellan M, Janu NC, Carmalt HL, Kennedy CW, Gillett DJ. Infiltrating lobular carcinoma—a comparison of diagnosis, management and outcome with infiltrating duct carcinoma. Breast. 2004;13(5):389–96.
26. Pestalozzi BC, Zahrieh D, Mallon E, Gusterson BA, Price KN, Gelber RD, et al. Distinct clinical and prognostic features of infiltrating lobular carcinoma of the breast: combined results of 15 International Breast Cancer Study Group clinical trials. J Clin Oncol. 2008;26(18):3006–14.
27. Mosiun JA, Idris MSB, Teoh LY, Teh MS, Chandran PA, See MH. Gastrointestinal tract metastasis presenting as intussusception in invasive lobular carcinoma of the breast: a case report. Int J Surg Case Rep. 2019;64:109–12.
28. Montagna E, Pirola S, Maisonneuve P, De Roberto G, Cancello G, Palazzo A, et al. Lobular metastatic breast cancer patients with gastrointestinal involvement: features and outcomes. Clin Breast Cancer. 2018;18(3):e401–5.
29. Viso Vidal D, Villanueva Pavón R, Jorquera PF. Linitis plastica of the colon due to metastases of invasive lobular breast carcinoma. Rev Esp Enferm Dig. 2019;111(4):326–8.
30. Ferlicot S, Vincent-Salomon A, Médioni J, Genin P, Rosty C, Sigal-Zafrani B, et al. Wide metastatic spreading in infiltrating lobular carcinoma of the breast. Eur J Cancer. 2004;40(3):336–41.
31. Briki R, Cherif O, Bannour B, Hidar S, Boughizane S, Khairi H. Uncommon metastases of invasive lobular breast cancer to the endometrium: a report of two cases and review of the literature. Pan Afr Med J. 2018;30:268.
32. Iorfida M, Maiorano E, Orvieto E, Maisonneuve P, Bottiglieri L, Rotmensz N, et al. Invasive lobular breast cancer: subtypes and outcome. Breast Cancer Res Treat. 2012;133(2):713–23.
33. Sastre-Garau X, Jouve M, Asselain B, Vincent-Salomon A, Beuzeboc P, Dorval T, et al. Infiltrating lobular carcinoma of the breast: clinicopathologic analysis of 975 cases with reference to data on conservative therapy and metastatic patterns. Cancer. 1996;77(1):113–20.
34. Butler D, Rosa M. Pleomorphic lobular carcinoma of the breast: a morphologically and clinically distinct variant of lobular carcinoma. Arch Pathol Lab Med. 2013;137(11):1688–92.
35. Eusebi V, Magalhaes F, Azzopardi JG. Pleomorphic lobular carcinoma of the breast: an aggressive tumor showing apocrine differentiation. Hum Pathol. 1992;23(6):655–62.
36. Widschwendter M, Jones PA. DNA methylation and breast carcinogenesis. Oncogene. 2002;21(35):5462–82.
37. Jovanovic J, Rønneberg JA, Tost J, Kristensen V. The epigenetics of breast cancer. Mol Oncol. 2010;4(3):242–54.
38. Droufakou S, Deshmane V, Roylance R, Hanby A, Tomlinson I, Hart IR. Multiple ways of silencing E-cadherin gene expression in lobular carcinoma of the breast. Int J Cancer. 2001;92(3):404–8.
39. Sarrió D, Moreno-Bueno G, Hardisson D, Sánchez-Estévez C, Guo M, Herman JG, et al. Epigenetic and genetic alterations of APC and CDH1 genes in lobular breast cancer: relationships with abnormal E-cadherin and catenin expression and microsatellite instability. Int J Cancer. 2003;106(2):208–15.
40. Caldeira JRF, Prando EC, Quevedo FC, Neto FAM, Rainho CA, Rogatto SR. CDH1promoter hypermethylation and E-cadherin protein expression in infiltrating breast cancer. BMC Cancer. 2006;6(1):48.
41. Zou D, Yoon H-S, Perez D, Weeks RJ, Guilford P, Humar B. Epigenetic silencing in non-neoplastic epithelia identifies E-cadherin (CDH1) as a target for chemoprevention of lobular neoplasia. J Pathol. 2009;218(2):265–72.
42. Fackler MJ, McVeigh M, Evron E, Garrett E, Mehrotra J, Polyak K, et al. DNA methylation of RASSF1A, HIN-1, RAR-beta, Cyclin D2 and Twist in in situ and invasive lobular breast carcinoma. Int J Cancer. 2003;107(6):970–5.
43. Seniski GG, Camargo AA, Ierardi DF, Ramos EAS, Grochoski M, Ribeiro ESF, et al. ADAM33 gene silencing by promoter hypermethylation as a molecular marker in breast invasive lobular carcinoma. BMC Cancer. 2009;9:80.

44. Lo P-K, Mehrotra J, D'Costa A, Fackler MJ, Garrett-Mayer E, Argani P, et al. Epigenetic suppression of secreted frizzled related protein 1 (SFRP1) expression in human breast cancer. Cancer Biol Ther. 2006;5(3):281–6.
45. Lehmann U, Celikkaya G, Hasemeier B, Länger F, Kreipe H. Promoter hypermethylation of the death-associated protein kinase gene in breast cancer is associated with the invasive lobular subtype. Cancer Res. 2002;62(22):6634–8.
46. Moelans CB, Vlug EJ, Ercan C, Bult P, Buerger H, Cserni G, et al. Methylation biomarkers for pleomorphic lobular breast cancer-a short report. Cell Oncol. 2015;38(5):397–405.
47. Bae YK, Brown A, Garrett E, Bornman D, Fackler MJ, Sukumar S, et al. Hypermethylation in histologically distinct classes of breast cancer. Clin Cancer Res. 2004;10(18 Pt 1):5998–6005.
48. Hall CA, Wang R, Miao J, Oliva E, Shen X, Wheeler T, et al. Hippo pathway effector Yap is an ovarian cancer oncogene. Can Res. 2010;70(21):8517–25.
49. Yin M, Zhang L. Hippo signaling: a hub of growth control, tumor suppression and pluripotency maintenance. J Genet Genom. 2011;38(10):471–81.
50. Li N, Xie C, Lu N. Crosstalk between Hippo signalling and miRNAs in tumour progression. FEBS J. 2017;284(7):1045–55.
51. Bailey CL, Kelly P, Casey PJ. Activation of Rap1 promotes prostate cancer metastasis. Can Res. 2009;69(12):4962–8.
52. McSherry EA, Brennan K, Hudson L, Hill AD, Hopkins AM. Breast cancer cell migration is regulated through junctional adhesion molecule-A-mediated activation of Rap1 GTPase. Breast Cancer Res. 2011;13(2):R31.
53. Ma X-L, Shen M-N, Hu B, Wang B-L, Yang W-J, Lv L-H, et al. CD73 promotes hepatocellular carcinoma progression and metastasis via activating PI3K/AKT signaling by inducing Rap1-mediated membrane localization of P110β and predicts poor prognosis. J Hematol Oncol. 2019;12(1):37.
54. Gozuacik D, Bialik S, Raveh T, Mitou G, Shohat G, Sabanay H, et al. DAP-kinase is a mediator of endoplasmic reticulum stress-induced caspase activation and autophagic cell death. Cell Death Differ. 2008;15(12):1875–86.
55. Tserga A, Michalopoulos NV, Levidou G, Korkolopoulou P, Zografos G, Patsouris E, et al. Association of aberrant DNA methylation with clinicopathological features in breast cancer. Oncol Rep. 2012;27(5):1630–8.
56. Thawani JP, Wang AC, Than KD, Lin C-Y, La Marca F, Park P. Bone morphogenetic proteins and cancer: review of the literature. Neurosurgery. 2010;66(2):233–46.
57. Du M, Su XM, Zhang T, Xing YJ. Aberrant promoter DNA methylation inhibits bone morphogenetic protein 2 expression and contributes to drug resistance in breast cancer. Mol Med Rep. 2014;10(2):1051–5.
58. Hung C-S, Wang S-C, Yen Y-T, Lee T-H, Wen W-C, Lin R-K. Hypermethylation of CCND2 in lung and breast cancer is a potential biomarker and drug target. Int J Mol Sci. 2018;19(10):3096.
59. Buffart TE, Overmeer RM, Steenbergen RD, Tijssen M, van Grieken NC, Snijders PJ, et al. MAL promoter hypermethylation as a novel prognostic marker in gastric cancer. Br J Cancer. 2008;99(11):1802–7.
60. Hu CY, Mohtat D, Yu Y, Ko Y-A, Shenoy N, Bhattacharya S, et al. Kidney cancer is characterized by aberrant methylation of tissue-specific enhancers that are prognostic for overall survival. Clin Cancer Res. 2014;20(16):4349–60.
61. Ellinger J, Bastian PJ, Jurgan T, Biermann K, Kahl P, Heukamp LC, et al. CpG island hypermethylation at multiple gene sites in diagnosis and prognosis of prostate cancer. Urology. 2008;71(1):161–7.
62. Guo W, Zhu L, Yu M, Zhu R, Chen Q, Wang Q. A five-DNA methylation signature act as a novel prognostic biomarker in patients with ovarian serous cystadenocarcinoma. Clin Epigenet. 2018;10(1):142.
63. Sailer V, Gevensleben H, Dietrich J, Goltz D, Kristiansen G, Bootz F, et al. Clinical performance validation of PITX2 DNA methylation as prognostic biomarker in patients with head and neck squamous cell carcinoma. PLoS ONE. 2017. https://doi.org/10.1371/journal.pone.0179412.
64. de Almeida BP, Apolónio JD, Binnie A, Castelo-Branco P. Roadmap of DNA methylation in breast cancer identifies novel prognostic biomarkers. BMC Cancer. 2019;19(1):219.
65. Xia B, Shan M, Wang J, Zhong Z, Geng J, He X, et al. Homeobox A11 hypermethylation indicates unfavorable prognosis in breast cancer. Oncotarget. 2017;8(6):9794.
66. Sheng X, Guo Y, Lu Y. Prognostic role of methylated GSTP1, p16, ESR1 and PITX2 in patients with breast cancer: a systematic meta-analysis under the guideline of PRISMA. Medicine. 2017;96(28):e7476.
67. Zhong Z, Shan M, Wang J, Liu T, Xia B, Niu M, et al. HOXD13 methylation status is a prognostic indicator in breast cancer. Int J Clin Exp Pathol. 2015;8(9):10716.
68. Martín-Sánchez E, Mendaza S, Ulazia-Garmendia A, Monreal-Santesteban I, Córdoba A, Vicente-García F, et al. CDH22 hypermethylation is an independent prognostic biomarker in breast cancer. Clin Epigenet. 2017;9(1):7.
69. Wu L, Wang F, Xu R, Zhang S, Peng X, Feng Y, et al. Promoter methylation of BRCA1 in the prognosis of breast cancer: a meta-analysis. Breast Cancer Res Treat. 2013;142(3):619–27.
70. Jiang Y, Cui L, Shen S, Ding L. The prognostic role of RASSF1A promoter methylation in breast cancer: a meta-analysis of published data. PLoS ONE. 2012;7(5):e36780.
71. Fleischer T, Klajic J, Aure MR, Louhimo R, Pladsen AV, Ottestad L, et al. DNA methylation signature (SAM40) identifies subgroups of the Luminal A breast cancer samples with distinct survival. Oncotarget. 2017;8(1):1074.
72. Stirzaker C, Zotenko E, Song JZ, Qu W, Nair SS, Locke WJ, et al. Methylome sequencing in triple-negative breast cancer reveals distinct methylation clusters with prognostic value. Nat Commun. 2015;6(1):1–11.
73. Debouki-Joudi S, Trifa F, Khabir A, Sellami-Boudawara T, Frikha M, Daoud J, et al. CpG methylation of APC promoter 1A in sporadic and familial breast cancer patients. Cancer Biomark. 2017;18(2):133–41.
74. He K, Zhang L, Long X. Quantitative assessment of the association between APC promoter methylation and breast cancer. Oncotarget. 2016;7(25):37920.
75. Brabender J, Usadel H, Danenberg KD, Metzger R, Schneider PM, Lord RV, et al. Adenomatous polyposis coli gene promoter hypermethylation in non-small cell lung cancer is associated with survival. Oncogene. 2001;20(27):3528–32.
76. Henrique R, Ribeiro FR, Fonseca D, Hoque MO, Carvalho AL, Costa VL, et al. High promoter methylation levels of APC predict poor prognosis in sextant biopsies from prostate cancer patients. Clin Cancer Res. 2007;13(20):6122–9.
77. Richiardi L, Fiano V, Vizzini L, De Marco L, Delsedime L, Akre O, et al. Promoter methylation in APC, RUNX3, and GSTP1 and mortality inprostate cancer patients. J Clin Oncol. 2009;27:3161–8.
78. Piqué L, de Paz AM, Piñeyro D, Martínez-Cardús A, de Moura MC, Llinàs-Arias P, et al. Epigenetic inactivation of the splicing RNA-binding protein CELF2 in human breast cancer. Oncogene. 2019;38(45):7106–12.
79. Milne RL, Fletcher AS, MacInnis RJ, Hodge AM, Hopkins AH, Bassett JK, et al. Cohort Profile: The Melbourne Collaborative Cohort Study (Health 2020). Int J Epidemiol. 2017;46(6):1757–i.
80. Wong EM, Joo JE, McLean CA, Baglietto L, English DR, Severi G, et al. Tools for translational epigenetic studies involving formalin-fixed paraffin-embedded human tissue: applying the Infinium HumanMethyation450 Beadchip assay to large population-based studies. BMC Res Notes. 2015;8:543.
81. Lakhani SR. WHO Classification of Tumours of the Breast: International Agency for Research on Cancer; 2012.
82. Wong EM, Joo JE, McLean CA, Baglietto L, English DR, Severi G, et al. Analysis of the breast cancer methylome using formalin-fixed paraffin-embedded tumour. Breast Cancer Res Treat. 2016;160(1):173–80.
83. Qin Y, Feng H, Chen M, Wu H, Zheng X. InfiniumPurify: an R package for estimating and accounting for tumor purity in cancer methylation research. Genes Dis. 2018;5(1):43–5.
84. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics. 2014;30(10):1363–9.
85. Fortin J-P, Labbe A, Lemire M, Zanke BW, Hudson TJ, Fertig EJ, et al. Functional normalization of 450k methylation array data improves replication in large cancer studies. Genome Biol. 2014;15(12):503.
86. Du P, Zhang X, Huang C-C, Jafari N, Kibbe WA, Hou L, et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinform. 2010;11:587.
87. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res. 2016;44(8):e71.
88. Peters TJ, Buckley MJ, Statham AL, Pidsley R, Samaras K, Lord R, et al. De novo identification of differentially methylated regions in the human genome. Epigenet Chromatin. 2015;8(1):6.

89. Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat Commun. 2019;10(1):1523.
90. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27–30.
91. Therneau T. A package for survival analysis in S. R package version 2.37-7. 2014.